

(19) 世界知的所有権機関
国際事務局(43) 国際公開日
2005 年 9 月 29 日 (29.09.2005)

PCT

(10) 国際公開番号
WO 2005/091146 A1

- (51) 国際特許分類: G06F 12/08
 (21) 国際出願番号: PCT/JP2005/004676
 (22) 国際出願日: 2005 年 3 月 16 日 (16.03.2005)
 (25) 国際出願の言語: 日本語
 (26) 国際公開の言語: 日本語
 (30) 優先権データ:
 特願2004-086174 2004 年 3 月 24 日 (24.03.2004) JP
 (71) 出願人 (米国を除く全ての指定国について): 松下電
 器産業株式会社 (MATSUSHITA ELECTRIC INDUS-
 TRIAL CO., LTD.) [JP/JP]; 〒5718501 大阪府門真市大
 字門真 1 0 0 6 番地 Osaka (JP).
 (72) 発明者; および
 (75) 発明者/出願人 (米国についてのみ): 中西 龍太

(NAKANISHI, Ryuta). 岡林 はづき (OKABAYASHI,
 Hazuki). 田中 哲也 (TANAKA, Tetsuya). 清原 督三
 (KIYOHARA, Tokuzou).

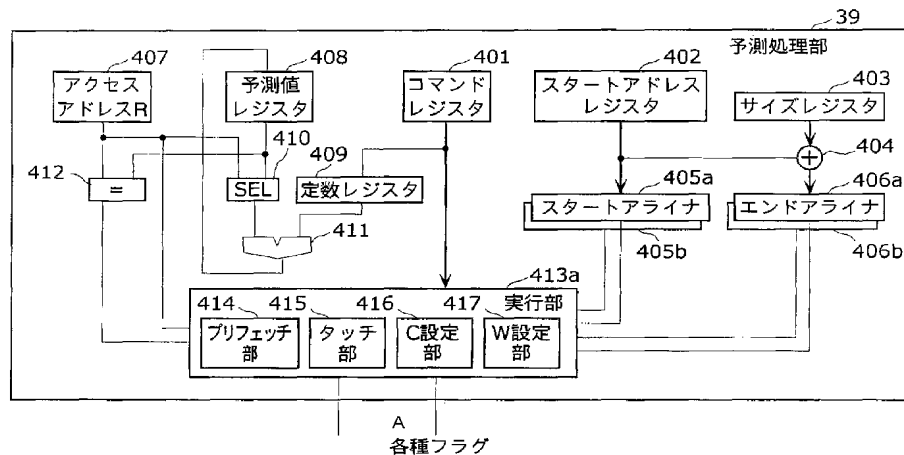
(74) 代理人: 新居 広守 (NII, Hiromori); 〒5320011 大阪府
 大阪市淀川区西中島 3 丁目 1 1 番 2 6 号 新大阪末広
 センタービル 3 F 新居国際特許事務所内 Osaka (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が
 可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR,
 BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM,
 DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU,
 ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS,
 LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA,
 NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE,
 SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG,
 US, UZ, VC, VN, YU, ZA, ZM, ZW.

[続葉有]

(54) Title: CACHE MEMORY AND CONTROL METHOD THEREOF

(54) 発明の名称: キャッシュメモリ及びその制御方法



39... PREDICTION UNIT
 407... ACCESS ADDRESS R
 408... PREDICTION VALUE REGISTER
 401... COMMAND REGISTER
 402... START ADDRESS REGISTER
 403... SIZE REGISTER
 409... CONSTANT REGISTER
 405a... START ALIGNER
 406a... END ALIGNER
 413a... EXECUTION UNIT
 414... PRE-FETCH UNIT
 415... TOUCH UNIT
 416... C SETTING UNIT
 417... W SETTING UNIT
 A...VARIOUS FLAGS

(57) Abstract: A cache memory includes a prediction unit (39) for predicting a line address to be pre-fetched according to the advance condition of memory access outputted from a processor. The prediction unit (39) has: a pre-fetch unit (414) for pre-fetching the predicted line address from the memory to the cache memory; and a touch unit (415) for setting the predicted line address as a tag in a cache entry to invalidate the valid flag without loading data from the memory to the cache memory.

(57) 要約: 本発明のキャッシュメモリは、プロセッサから出力されるメモリアクセスの進行状況に基づいて次にプ
 リフェッチすべきラインアドレスを予測する予測処理部 39 を有し、予測処理部 39 は、メモリからキャッシュメモ

[続葉有]

WO 2005/091146 A1



(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

2文字コード及び他の略語については、定期発行される各*PCT*ガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

添付公開書類:

— 国際調査報告書

りに予測されたラインアドレスのデータをプリフェッチするプリフェッチ部414と、メモリからキャッシュメモリにデータをロードすることなく、予測されたラインアドレスをタグとしてキャッシュエントリーに設定し、バリッドフラグを有効にするタッチ部415とを備える。

明 細 書

キャッシュメモリ及びその制御方法

技術分野

- [0001] 本発明は、プロセッサのメモリアクセスを高速化するためのキャッシュメモリ及びその制御方法に関する。

背景技術

- [0002] 近年のマイクロプロセッサでは、例えば、SRAM(Static Random Access Memory)等から成る小容量で高速なキャッシュメモリをマイクロプロセッサの内部、もしくはその近傍に配置し、データの一部をキャッシュメモリに記憶することによって、マイクロプロセッサのメモリアクセスを高速化させている。
- [0003] キャッシュの効率向上(ヒット率向上、キャッシュミスレイテンシ低減)のため、キャッシュミスが発生する前に、近い未来に使用するデータを予めキャッシュにフィルする技術としてプリロード(又はプリフェッチ)がある(例えば、特許文献1)。
- [0004] 従来のプリフェッチ技術では、プリフェッチ命令により指定したアドレスを含むラインをキャッシュにロードしている。これにより、キャッシュミスの低減を図っている。

特許文献1:特開平7-295882号公報

発明の開示

発明が解決しようとする課題

- [0005] しかしながら、上記従来技術によれば、例えば、ソフトウェアによって、ループの外側で一括してプリフェッチを行った場合、ループで必要なデータ領域をあらかじめすべてキャッシュ上に確保することになるため、キャッシュの容量が小さい場合には、それ以外の必要なデータがキャッシュから追い出され、キャッシュミスが発生する。また、データのキャッシング終了、無効化等をループの外側で一括して行う場合は、ループを抜けるまでこれらの動作によるキャッシュの開放が行われないため、キャッシュの容量が不足し、キャッシュミスが発生する。
- [0006] また、ループ中にソフトウェアによってキャッシュ操作を行う命令を挿入した場合、キャッシュ操作を行うアドレスをループの中でソフトウェアによって管理する必要がある。

つまり、キャッシュ操作を行う命令をループの中に記述する必要があるため、性能の劣化が発生する。

- [0007] さらに、メモリへのアクセスの状況をハードウェアによって監視し、ハードウェアによって自動的にキャッシュ操作をする場合、正確な予測が行えないと無駄な転送が発生する、あるいは、ソフトウェアからの正確な情報がないとキャッシュ上のデータと外部メモリ上のデータの整合性が取ることができないため、ハードウェアの予測によってこれらの操作を行うことは困難である。

課題を解決するための手段

- [0008] 本発明の目的は、プロセッサからあらかじめ設定された情報を元に、ハードウェアがプロセッサのプログラム動作の状況を監視し、同期を取りつつ適切なタイミングでキャッシュに対して操作を行うキャッシュメモリシステムを提供することにある。
- [0009] 上記目的を達成するため本発明のキャッシュメモリシステムは、プロセッサの状態に関する条件を生成する条件生成手段と、現在のプロセッサの状態が前記条件を満たすかどうかを判定する判定手段と、操作対象となるアドレスを生成するアドレス生成手段と、前記判定手段が条件を満たすと判定したときに前記アドレス生成手段によって生成されたアドレスを用いてキャッシュを操作する操作手段とを備える。
- [0010] ここで、前記条件生成手段は、前記判定手段が条件を満たすと判定した場合に新たな条件を生成するようにしてもよい。
- [0011] この構成によれば、プロセッサの動作状態が条件を満たすようになったときにキャッシュを操作するので、プロセッサの動作の進行状況に同期してキャッシュメモリを操作することができる。またソフトウェア的に実現しないのでプロセッサに負荷をかけることなくキャッシュメモリを効率よく性能劣化を招くことなく動作させることができる。
- [0012] ここで、前記条件生成手段は、プロセッサ内の特定レジスタの値に関する条件を生成するようにしてもよい。前記特定レジスタはプログラムカウンタであってもよい。
- [0013] この構成によれば、メモリアクセスアドレスや、プログラムフェッチアドレスを前記条件として、プロセッサの状態を監視することができる。
- [0014] ここで、前記条件生成手段は、特定のアドレス範囲内へのメモリアクセスおよび特定のアドレス範囲外へのメモリアクセスの何れかを前記条件として生成するようにしても

よい。

[0015] また、前記条件生成手段は、プロセッサが特定命令を実行することを前記条件として生成するようにしてもよい。

[0016] ここで、前記条件生成手段は、現在の条件に特定の演算を施すことによって前記新たな条件を生成するようにしてもよい。

[0017] また、前記条件生成手段はメモリアクセスアドレスを条件として生成し、前記判定手段が条件を満たすと判定した場合に現在の条件に定数を加算することによって前記新たな条件を生成するようにしてもよい。

[0018] ここで、前記定数は、プロセッサにより実行されるポストインクリメント付きロード／ストア命令におけるインクリメント値またはデクリメント値、およびプロセッサにより実行される2回のロード／ストア命令におけるアドレスの差分値の何れかであるようにしてもよい。

[0019] ここで、前記条件生成手段は複数の条件を生成し、前記判定手段は、複数の条件のすべてを満たすかどうかを判定するようにしてもよい。

[0020] また、前記条件生成手段は複数の条件を生成し、前記判定手段は、複数の条件の何れかを満たすかどうかを判定するようにしてもよい。

[0021] ここで、前記操作手段は、前記判定手段が条件を満たすと判定したときに、前記アドレス生成手段により生成されたアドレスに対応するデータがキャッシュに格納されているかどうかを判定するデータ判定手段と、格納されていないと判定された場合に、キャッシュメモリ中のラインを選択する選択手段と、前記選択されたラインが有効でデータならライトバックを行うライトバック手段と、前記アドレスに対応するデータをメモリからライトバック後の選択されたラインへ転送する転送手段と、前記アドレスをタグとして前記選択されたラインへ登録する登録手段とを備えるよう構成してもよい。

[0022] この構成によれば、キャッシュメモリのプリフェッチを、プロセッサの動作状況を監視して同期を取りつつ適切なタイミングで行うことができる。

[0023] ここで、前記操作手段は、前記判定手段が条件を満たすと判定したときに、前記アドレス生成手段により生成されたアドレスに対応するデータがキャッシュに格納されているかどうかを判定するデータ判定手段と、格納されていないと判定された場合に、

キャッシュメモリ中のラインを選択する選択手段と、選択されたラインが有効でデータであれば、ライトバックを行うライトバック手段と、メモリから選択されたラインへデータを転送することなく、前記生成したアドレスをタグとして選択されたラインへ登録する登録手段とを備える構成としてもよい。

[0024] この構成によれば、キャッシュメモリのラインにデータを転送することなく確保すること（タッチと呼ぶ）を、プロセッサの動作状況を監視して同期を取りつつ適切なタイミングで行うことができる。

[0025] ここで、前記操作手段は、前記判定手段が条件を満たすと判定したときに、前記アドレス生成手段により生成されたアドレスに対応するデータがキャッシュに格納されているかどうかを判定するデータ判定手段と、格納されていると判定された場合に、キャッシュメモリ中の格納先のラインを選択する選択手段と、選択されたラインが有効でかつデータであればライトバックを行うライトバック手段とを備える構成としてもよい。

[0026] この構成によれば、キャッシュメモリのラインデータのライトバックを（キャッシング終了）を、プロセッサの動作状況を監視して同期を取りつつ適切なタイミングで行うことができる。

[0027] ここで、前記操作手段は、前記判定手段が条件を満たすと判定したときに、前記アドレス生成手段により生成されたアドレスに対応するデータがキャッシュに格納されているかどうかを判定するデータ判定手段と、格納されていると判定された場合に、キャッシュメモリ中の格納先のラインを選択する選択手段と、選択されたラインを無効化する無効化手段とを備える構成としてもよい。

[0028] この構成によれば、キャッシュメモリのラインの無効化を、プロセッサの動作状況を監視して同期を取りつつ適切なタイミングで行うことができる。

[0029] ここで、前記操作手段は、前記判定手段が条件を満たすと判定したときに、前記アドレス生成手段により生成されたアドレスに対応するデータがキャッシュに格納されているかどうかを判定するデータ判定手段と、格納されていると判定された場合に、キャッシュメモリ中の格納先のラインを選択する選択手段と、ラインのアクセス順序を示す順序情報に対して、選択されたラインのアクセス順序を変更する変更手段とを備える構成としてもよい。

- [0030] この構成によれば、キャッシュメモリのラインのアクセス順序情報を、プロセッサの動作状況を監視して同期を取りつつ適切なタイミングで行うことができる。これによりいわゆるLRUによるキャッシュのリプレース順序を早めたり遅くしたりすることができる。
- [0031] ここで、前記条件生成手段により前記条件としてメモリアドレスを生成し、前記操作手段は、さらに、前記条件生成手段により生成されたメモリアドレスがラインの途中を指す場合に、当該ラインの先頭、次のラインの先頭および前のラインの先頭の何れかを指すように調整することによりアドレスを生成する調整手段を備える構成としてもよい。
- [0032] この構成によれば、プロセッサがアドレス昇順にアクセスする場合も、アドレス高順にアクセスする場合も、次に必要なラインのアドレスを適切に算出することができる。
- [0033] また、本発明のキャッシュメモリの制御方法についても上記と同様の手段、作用を有する。
- [0034] 本発明のキャッシュメモリによれば、プロセッサの動作の進行状況に同期してキャッシュメモリを操作することができる。またソフトウェア的に実現しないのでプロセッサに負荷をかけることなくキャッシュメモリを効率よく性能劣化を招くことなく動作させることができる。
- [0035] 例えば、プロセッサがポストインクリメント付きロード命令によってシーケンシャルにメモリアクセスする場合には、効率よくプリフェッチすることができる。また、プロセッサがポストインクリメント付きストア命令によってシーケンシャルにデータを書き込む場合には、メモリからキャッシュメモリにデータをロードするペナルティを削除することができ、さらに効率よくキャッシュエントリーをタッチ(確保)することができる。
- [0036] また、プロセッサがポストインクリメント付きストア命令によってシーケンシャルにデータを書き込む場合に、プロセッサのストア命令の進行状況に応じて、ストアが完了したラインを予測し、予測したラインにキャッシング終了属性や無効化属性を設定するので、プロセッサでは、キャッシュメモリのラインサイズやライン境界を認識しなくても、ライトバックすべきラインをキャッシュメモリにおいて予測し、効率よくクリーニング(ライトバック)や、キャッシュエントリーの開放(無効化)を行うことができる。
- [0037] さらに、プロセッサがポストインクリメント付きロード命令によってシーケンシャルにデ

ータを読み出す場合に、読み出しが完了したラインデータを保持するキャッシュエントリーにリプレース属性が設定され、真っ先にリプレース対象として選択されるので、アクセス頻度の低いデータがキャッシュメモリに残ることによるキャッシュミス誘発を低減することができる。

発明の効果

- [0038] 本発明のキャッシュメモリによれば、プロセッサの動作の進行状況に同期してキャッシュメモリを操作することができる。またソフトウェア的に実現しないのでプロセッサに負荷をかけることなくキャッシュメモリを効率よく性能劣化を招くことなく動作させることができる。また、メモリアクセスアドレスや、プログラムフェッチアドレスを前記条件として、プロセッサの状態を監視することができる。
- [0039] たとえば、キャッシュメモリのラインにデータを転送するとなく確保すること(タッチと呼ぶ)を、プロセッサの動作状況を監視して同期を取りつつ適切なタイミングで行うことができる。
- [0040] また、キャッシュメモリのラインデータのライトバックを(キャッシング終了)を、プロセッサの動作状況を監視して同期を取りつつ適切なタイミングで行うことができる。キャッシュメモリのラインの無効化を、プロセッサの動作状況を監視して同期を取りつつ適切なタイミングで行うことができる。
- [0041] さらに、キャッシュメモリのラインのアクセス順序情報を、プロセッサの動作状況を監視して同期を取りつつ適切なタイミングで行うことができる。これによりいわゆるLRUによるキャッシュのリプレース順序を早めたり遅くしたりすることができる。
- [0042] また、プロセッサがアドレス昇順にアクセスする場合も、アドレス高順にアクセスする場合も、次に必要なラインのアドレスを適切に算出することができる。

図面の簡単な説明

- [0043] [図1]図1は、本発明の実施の形態1におけるプロセッサ、キャッシュメモリ、メモリを含むシステムの概略構成を示すブロック図である。
- [図2]図2は、キャッシュメモリの構成例を示すブロック図である。
- [図3]図3は、予測処理部の構成例を示すブロック図である。
- [図4]図4(a)は、スタートアドレスレジスタにスタートアドレスを書き込む命令の一例を

示す。図4(b)は、サイズレジスタにサイズを書き込む命令の一例を示す。図4(c)は、コマンドレジスタにコマンドを書き込む命令の一例を示す。図4(d)は、コマンドの一例を示す。

[図5]図5は、スタートアライナ及びエンドアライナの説明図である。

[図6]図6(a)は、プリフェッチ部による予測プリフェッチの説明図である。図6(b)は、タッチ部による予測タッチの説明図である。

[図7]図7は、プリフェッチ部における予測プリフェッチ処理の一例を示すフローチャートである。

[図8]図8は、タッチ部における予測タッチ処理の一例を示すフローチャートである。

[図9]図9は、本発明の実施の形態2におけるキャッシュメモリの構成を示すブロック図である。

[図10]図10は、キャッシュエントリーのビット構成を示す。

[図11]図11は、制御部による使用フラグの更新例を示す。

[図12]図12(a)は、ウィークフラグが存在しない場合にキャッシュエントリーがリプレイスされる様子を示す図である。図12(b)リプレイス処理におけるウィークフラグWの役割を示す説明図である。

[図13]図13は、予測処理部の構成を示すブロック図である。

[図14]図14(a)は、C設定部によるCフラグ設定処理の説明図である。図14(b)は、W設定部によるWフラグ設定処理の説明図である。

[図15]図15は、Cフラグ設定部におけるCフラグ設定処理の一例を示すフローチャートである。

[図16]図16は、Wフラグ設定部におけるWフラグ設定処理の一例を示すフローチャートである。

[図17]図17は、制御部におけるクリーニング処理の一例を示すフローチャートである。

[図18]図18は、制御部におけるUフラグ更新処理を示すフローチャートである。

[図19]図19は、制御部におけるリプレイス処理を示すフローチャートである。

符号の説明

- [0044]
- 1 プロセッサ
 - 2 メモリ
 - 3 キャッシュメモリ
 - 20 アドレスレジスタ
 - 21 メモリI/F
 - 30 デコーダ
 - 31a〜31d ウエイ
 - 32a〜32d 比較器
 - 33a〜33d アンド回路
 - 34 オア回路
 - 35 セレクタ
 - 36 セレクタ
 - 37 デマルチプレクサ
 - 38 制御部
 - 39 予測処理部
 - 131a ウエイ
 - 131a〜131d ウエイ
 - 131b〜131d ウエイ
 - 138 制御部
 - 139 予測処理部
 - 401 コマンドレジスタ
 - 402 スタートアドレスレジスタ
 - 403 サイズレジスタ
 - 404 加算器
 - 405a、405b スタートアライナ
 - 406a、406b エンドアライナ
 - 407 アクセスアドレスレジスタ
 - 408 予測値レジスタ

- 409 定数レジスタ
- 410 セレクタ
- 411 加算器
- 412 比較器
- 413 実行部
- 413a 実行部
- 414 プリフェッチ部
- 415 タッチ部
- 416 C設定部
- 417 W設定部

発明を実施するための最良の形態

[0045] (実施の形態1)

<全体構成>

図1は、本発明の実施の形態1におけるプロセッサ1、キャッシュメモリ3、メモリ2を含むシステムの概略構成を示すブロック図である。同図のように、本発明のキャッシュメモリ3は、プロセッサ1およびメモリ2を有するシステムに備えられる。

[0046] キャッシュメモリ3は、プロセッサ1から指定された条件に従って、予測プリフェッチを行うよう構成されている。予測プリフェッチとは、プロセッサ1によるメモリアクセスの進行状況に基づいて次にプリフェッチすべきラインアドレスを予測し、予測したラインアドレスのデータをキャッシュメモリにプリフェッチすることをいう。また、プロセッサ1から指定される条件というのは、アドレス範囲、シーケンシャルアクセスする場合のアドレスのインクリメント値またはデクリメント値、またはアクセス方向(アドレス昇順または降順)等である。予測プリフェッチは、プロセッサがシーケンシャルにメモリアクセスする場合に適している。

[0047] 加えて、キャッシュメモリ3は、プロセッサ1から指定された条件に従って、予測タッチを行うよう構成されている。予測タッチとは、プロセッサ1によるメモリアクセスの進行状況に基づいて次にタッチすべきラインアドレスを予測し、予測したラインアドレスのデータをロードすることなくセットキャッシュエントリーにバリッドフラグとタグとを設定する

ことにより確保することをいう。予測タッチは、配列データなどの演算結果をメモリに順次格納する場合などに適している。

[0048] <キャッシュメモリの構成>

以下、キャッシュメモリ3の具体例として、4ウェイ・セット・アソシエティブ方式のキャッシュメモリに本発明を適用した場合の構成について説明する。

[0049] 図2は、キャッシュメモリ3の構成例を示すブロック図である。同図のように、キャッシュメモリ3は、アドレスレジスタ20、メモリI/F21、デコーダ30、4つのウェイ31a〜31d(以下ウェイ0〜3と略す)、4つの比較器32a〜32d、4つのアンド回路33a〜33d、オア回路34、セクタ35、36、デマルチプレクサ37、制御部38を備える。

[0050] アドレスレジスタ20は、メモリ2へのアクセスアドレスを保持するレジスタである。このアクセスアドレスは32ビットであるものとする。同図に示すように、アクセスアドレスは、最上位ビットから順に、21ビットのタグアドレス、4ビットのセットインデックス(図中のSI)、5ビットのワードインデックス(図中のWI)を含む。ここで、タグアドレスはウェイにマッピングされるメモリ中の領域(そのサイズはセット数×ブロックである)を指す。この領域のサイズは、タグアドレスよりも下位のアドレスビット(A10〜A0)で定まるサイズつまり2kバイトであり、1つのウェイのサイズでもある。セットインデックス(SI)はウェイ0〜3に跨る複数セットの1つを指す。このセット数は、セットインデックスが4ビットなので16セットある。タグアドレスおよびセットインデックスで特定されるキャッシュエントリは、リプレース単位であり、キャッシュメモリに格納されている場合はラインデータ又はラインと呼ばれる。ラインデータのサイズは、セットインデックスよりも下位のアドレスビットで定まるサイズつまり128バイトである。1ワードを4バイトとすると、1ラインデータは32ワードである。ワードインデックス(WI)は、ラインデータを構成する複数ワード中の1ワードを指す。アドレスレジスタ20中の最下位2ビット(A1、A0)は、ワードアクセス時には無視される。

[0051] メモリI/F21は、キャッシュメモリ3からメモリ2へのデータのライトバックや、メモリ2からキャッシュメモリ3へのデータのロード等、キャッシュメモリ3からメモリ2をアクセスするためのI/Fである。

[0052] デコーダ30は、セットインデックスの4ビットをデコードし、4つのウェイ0〜3に跨る1

6セット中の1つを選択する。

- [0053] 4つのウェイ0〜3は、同じ構成を有数する4つのウェイであり、 $4 \times 2k$ バイトの容量を有する。各ウェイは、16個のキャッシュエントリーを有する。1つのキャッシュエントリーは、バリッドフラグV、21ビットのタグ、128バイトのラインデータ、ダーティフラグDを有する。タグは21ビットのタグアドレスのコピーである。ラインデータは、タグアドレスおよびセットインデックスにより特定されるブロック中の128バイトデータのコピーである。バリッドフラグVは、当該キャッシュエントリーのデータが有効か否かを示す。ダーティフラグDは、当該キャッシュエントリーにプロセッサから書き込みがあったか否か、つまりサブライン中にキャッシュされたデータが存在するが書き込みによりメモリ中のデータと異なるためメモリに書き戻すことが必要か否かを示す。
- [0054] 比較器32aは、アドレスレジスタ20中のタグアドレスと、セットインデックスにより選択されたセットに含まれる4つのタグ中のウェイ0のタグとが一致するか否かを比較する。比較器32b〜32cについても、ウェイ31b〜31dに対応すること以外は同様である。
- [0055] アンド回路33aは、バリッドフラグと比較器32aの比較結果とが一致するか否かを比較する。この比較結果をh0とする。比較結果h0が1である場合は、アドレスレジスタ20中のタグアドレスおよびセットインデックスに対応するラインデータが存在すること、つまりウェイ0においてヒットしたことを意味する。比較結果h0が0である場合は、ミスヒットしたことを意味する。アンド回路33b〜33dについても、ウェイ31b〜31dに対応すること以外は同様である。その比較結果h1〜h3は、ウェイ1〜3でヒットしたかミスしたかを意味する。
- [0056] オア回路34は、比較結果h0〜h3のオアをとる。このオアの結果をhitとする。hitは、キャッシュメモリにヒットしたか否かを示す。
- [0057] セレクタ35は、選択されたセットにおけるウェイ0〜3のラインデータのうち、ヒットしたウェイのラインデータを選択する。
- [0058] セレクタ36は、セレクタ35により選択された32ワードのラインデータのうち、ワードインデックスに示される1ワードを選択する。
- [0059] デマルチプレクサ37は、キャッシュエントリーにデータを書き込む際に、ウェイ0〜3の1つに書き込みデータを出力する。この書き込みデータはワード単位でよい。

[0060] 制御部38は、予測処理部39を含み、キャッシュメモリ3の全体の制御を行う。予測処理部39は、主としてプリフェッチの制御を行う。

[0061] <予測処理部の構成>

図3は、予測処理部39の構成例を示すブロック図である。同図のように予測処理部39は、コマンドレジスタ401、スタートアドレスレジスタ402、サイズレジスタ403、加算器404、スタートアライナ405a、405b、エンドアライナ406a、406b、アクセスアドレスレジスタ407、予測値レジスタ408、定数レジスタ409、セクタ410、加算器411、比較器412、実行部413を備える。

[0062] コマンドレジスタ401は、プロセッサ1から直接アクセス可能なレジスタであり、プロセッサ1により書き込まれたコマンドを保持する。図4(c)に、コマンドレジスタ401にコマンドを書き込む命令の一例を示す。この命令は、通常の転送命令(mov命令)であり、ソースオペランドとしてコマンドを、デスティネーションオペランドとしてコマンドレジスタ(CR)401を指定している。図4(d)に、コマンドフォーマットの一例を示す。このコマンドフォーマットは、コマンド内容と定数とからなる。ここで、コマンド内容は、予測プリフェッチコマンドと、予測タッチコマンドとの何れかを表す。定数は、プロセッサ1が、シーケンシャルにメモリアクセスする場合のメモリアクセスアドレスのインクリメント値またはデクリメント値を表し、例えば、+4、+8、-4、-8等である。なお、コマンド中の定数の代わりに、アクセス方向(アドレス昇順かアドレス降順か)と絶対値(連続アクセスする場合のアドレスの差分)とを含むようにしてもよい。

[0063] スタートアドレスレジスタ402は、プロセッサ1から直接アクセス可能なレジスタであり、プロセッサ1により書き込まれたスタートアドレスを保持する。このスタートアドレスはCフラグ(キャッシングを終了してよいか否かを示すクリーニングフラグ)を設定すべきアドレス範囲の開始位置を示す。図4(a)に、スタートアドレスレジスタ402にスタートアドレスを書き込む命令の一例を示す。この命令も、図4(c)と同様に通常の転送命令(mov命令)である。

[0064] サイズレジスタ403は、プロセッサ1から直接アクセス可能なレジスタであり、プロセッサ1により書き込まれたサイズを保持する。このサイズは、スタートアドレスからのアドレス範囲を示す。図4(b)に、サイズレジスタ403にサイズを書き込む命令の一例を示す。

す。この命令も、図4(c)と同様に通常の転送命令(mov命令)である。なお、サイズの単位は、バイト数であっても、ライン数(キャッシュエントリー数)であってもよく、予め定められた単位であればよい。

- [0065] 加算器404は、スタートアドレスレジスタ402に保持されたスタートアドレスとサイズレジスタ403に保持されたサイズとを加算する。加算結果は、アドレス範囲の終了位置を指すエンドアドレスである。加算器404は、サイズがバイト数指定の場合はバイトアドレスとして加算し、サイズがライン数指定の場合はラインアドレスとして加算すればよい。
- [0066] スタートアライナ405a、405bは、スタートアドレスをライン境界の位置に調整する。スタートアライナ405aはエンドアドレスの方向に、405bはエンドアドレスとは反対の方向に調整する。この調整によりプロセッサ1はラインサイズ及びライン境界とは無関係に任意のアドレスをスタートアドレスとして指定することができる。
- [0067] エンドアライナ406a、406bは、エンドアドレスをライン境界の位置に調整する。エンドアライナ406aはスタートアドレスの方向に、406bはスタートアドレスとは反対の方向に調整する。この調整によりプロセッサ1はラインサイズ及びライン境界とは無関係に任意の大きさを上記サイズとして指定することができる。
- [0068] 図5に、スタートアライナ405a、405b及びエンドアライナ406a、406bの説明図を示す。同図において、プロセッサ1から指定されたスタートアドレスはラインNの途中の任意の位置を指す。スタートアライナ405aは、次のライン(N+1)の先頭を指すよう調整し、調整後のアドレスをアラインスタートアドレスaとして出力する。スタートアライナ405bは、スタートアドレスのデータを含むラインNの先頭を指すよう調整し、調整後のアドレスをアラインスタートアドレスbとして出力する。アラインスタートアドレスが指すラインをスタートラインと呼ぶ。
- [0069] また、エンドアドレスはラインMの途中の任意の位置を指す。エンドアライナ406aは、直前のライン(M-1)の先頭を指すよう調整し、調整後のアドレスをアラインエンドアドレスaとして出力する。エンドアライナ406bは、エンドアドレスのデータを含むラインMの先頭を指すよう調整し、調整後のアドレスをアラインエンドアドレスbとして出力する。アラインエンドアドレスが指すラインをエンドラインと呼ぶ。

- [0070] 同図のように、スタートアライナ405a及びエンドアライナ406aはライン単位で内側アラインを行う。スタートアライナ405b及びエンドアライナ406bはライン単位で外側アラインを行う。さらに、ライン単位の外側アラインの後、さらに、サブライン単位の外側アラインと内側アラインが可能である。
- [0071] アクセスアドレスレジスタ407は、プロセッサ1からのメモリアクセスアドレスを保持する。
- [0072] 予測値レジスタ408は、初期値として、アクセスアドレスレジスタ407のメモリアクセスアドレスと定数レジスタ409の定数とを加算した値を予測値として保持し、以降、プロセッサ1がメモリアクセスを実行したときに、アクセスアドレスレジスタ407のメモリアクセスアドレスと予測値とが一致していれば、予測値と定数レジスタ409の定数とを加算した値を新たな予測値として更新し、一致していなければ、新たな初期値に更新する。
- [0073] 定数レジスタ409は、プロセッサ1によるメモリアクセスアドレスのインクリメント値またはデクリメント値を保持する。このインクリメント値(またはデクリメント値)は、プロセッサ1がメモリをシーケンシャルにアクセスする場合のポストインクリメント付きロード/ストア命令のインクリメント値(またはデクリメント値)であり、例えば、図4(d)に示したコマンド中の定数として指定される。
- [0074] セレクタ410は、アクセスアドレスレジスタ407のメモリアクセスアドレスと予測値レジスタ408の予測値とが一致していれば、予測値レジスタ408を選択し、一致していなければ、アクセスアドレスレジスタ407を選択する。
- [0075] 加算器411は、セレクタ410に選択された予測値またはメモリアクセスアドレスと、定数レジスタ409の定数とを加算する。加算後の値は、新たな予測値又は新たな初期値として予測値レジスタ408に保持される。
- [0076] 比較器412は、アクセスアドレスレジスタ407のメモリアクセスアドレスと予測値レジスタ408の予測値とが一致するか否かを判定する。
- [0077] 実行部413は、プリフェッチ部414とタッチ部415とを備える。
- プリフェッチ部414は、プロセッサのロード命令の進行状況に応じて次にロードされるライン推定し、推定したラインをプリフェッチする予測プリフェッチを行う。

[0078] タッチ部415は、プロセッサのストア命令の進行状況に応じて次にストアされるライン推定し、推定したラインを保持するためのキャッシュエントリーを、メモリからデータをロードすることなく確保する(V=1の設定とタグの設定とを行う)。

[0079] <予測プリフェッチ>

図6(a)は、プリフェッチ部414による予測プリフェッチの説明図である。同図(a)において、ラインN、N+1、・・・、N+nはアドレスが連続するラインを示す。スタートアドレスレジスタ402及びサイズレジスタ403により定まるスタートライン、エンドラインは、それぞれラインN+1、ラインN+nであるものとする。また、LD(1)、LD(2)、・・・は、ロード命令によるアクセス位置を示している。

[0080] プリフェッチ部414は、最初のロード命令LD(1)の実行時に、そのメモリアクセスアドレスに定数を加算した値を初期値として予測値レジスタ408に保持させる。2回目のロード命令LD(2)の実行時に、そのメモリアクセスアドレスと予測値レジスタ408の予測値とが一致した場合、プリフェッチ部414は、ポストインクリメント付きロード命令によるシーケンシャルアクセスであるものと推定して、次のラインN+1をプリフェッチする。もし、次のラインがキャッシュメモリに格納されている場合は、なにもしない。

[0081] プロセッサ1によってポストインクリメント付きロード命令によりシーケンシャルアクセスがなされた場合、予測値はメモリアクセス毎に一致していくことになる。

[0082] このようにして、予測値レジスタ408の予測値がラインNからラインN+1のアドレスにまで更新され、メモリアクセスアドレスと一致したとき、プリフェッチ部414は、ラインN+2をプリフェッチする。プリフェッチ部414は、この予測プリフェッチをスタートラインからエンドラインまで実行し、エンドラインのプリフェッチが完了すると予測プリフェッチを終了する。

[0083] <予測タッチ>

図6(b)は、タッチ部415による予測タッチの説明図である。図6(a)と比較して、プロセッサ1がストア命令によりシーケンシャルにアクセスする場合に、そのメモリアクセスアドレスと予測値レジスタ408の予測値とが一致した場合に、ポストインクリメント付きストア命令によるシーケンシャルアクセスと推定し、次のラインN+1に対応するキャッシュエントリーを確保する点が異なる。つまり、プリフェッチの代わりに、キャッシュエ

ントリーにメモリデータをロードしないでタグ及びバリッドフラグを1に設定する点が異なっている。これ以外は予測プリフェッチと同様なので説明を省略する。

[0084] <予測プリフェッチ処理>

図7は、プリフェッチ部414における予測プリフェッチ処理の一例を示すフローチャートである。

同図において、プリフェッチ部414は、コマンドレジスタ401に予測プリフェッチコマンドが保持されていて(S41)、プロセッサがロード命令を実行したとき(S42)、当該メモリアクセスアドレスと定数とを加算した値を初期値として予測値レジスタ408に設定する(S43)。この最初のロード命令のメモリアクセスアドレスと予測値レジスタの予測値とは当然一致しない。予測値レジスタ408はクリアされているかランダムな値を保持しているからである。

[0085] さらに、プロセッサがロード命令を実行し(S44)、かつメモリアクセスアドレスと予測値レジスタの予測値とが一致する場合には(S45)、プリフェッチ部414は、次ラインのラインアドレスを算出し(S46)、算出したラインアドレスがスタートラインからエンドラインまでのアドレス範囲に属していて(S47)、次ラインがキャッシュメモリにエントリーされていないならば(S48)、次ラインをプリフェッチする(S49)。

[0086] このプリフェッチにおいて、プリフェッチ部414は、LRU方式でリプレース対象のウェイを選択し(S401)、当該ウェイのキャッシュエントリーがダーティであればライトバックし(S402、S403)、次ラインのデータを当該キャッシュエントリーにリフィル(プリフェッチ)する(S404)。

[0087] さらに、プリフェッチ部414は、プリフェッチしたラインがエンドラインであれば(S50)予測プリフェッチ処理を終了する。

[0088] また、上記S45において、メモリアクセスアドレスと予測値レジスタの予測値とが一致しない場合には、メモリアクセスアドレスに定数を加算した値が新たな予測値として予測値レジスタ408に設定される。また、一致しない場合は、なにもしない(メモリアクセス待ちの状態)。

[0089] このように、プリフェッチ部414は、予測プリフェッチにおいて、プロセッサのロード命令の進行状況に応じて次のラインを推定し、推定したラインをプリフェッチするので、

プロセッサ1では、キャッシュメモリのラインサイズやライン境界を認識する必要がなく、つまり、プリフェッチ用のアドレスをプロセッサ1にて管理する必要がない。キャッシュメモリにおいて推定に基づいて効率よくプリフェッチを行うことができる。

[0090] <予測タッチ処理>

図8は、タッチ部415における予測タッチ処理の一例を示すフローチャートである。同図は、図7の予測プリフェッチ処理と比較して、S41、S42、S44、S49、S404の各ステップの代わりにS41a、S42a、S44a、S49a、S404aを有する点が異なっている。これ以外は同じなので異なる点について説明する。

[0091] タッチ部415は、S41aにおいて、予測タッチコマンドがコマンドレジスタ401に保持されているかを判定し、S42a、S44aにおいてストア命令が実行されたか否かを判定する。また、タッチ部415は、S49aにおいてプリフェッチする代わりにタッチする。すなわち、S404aにおいて、リプレース対象のキャッシュエントリーメモリデータをロードしないでタグ及びバリッドフラグを1に設定する。

[0092] このように、タッチ部415は、プロセッサのストア命令の進行状況に応じて次のラインを推定し、推定したラインをタッチする(確保する)ので、プロセッサ1では、キャッシュメモリのラインサイズやライン境界を認識する必要がなく、つまりプリフェッチ用のアドレスをプロセッサ1にて管理する必要がない。キャッシュメモリにおいて推定に基づいて効率よくキャッシュエントリーをタッチすることができる。

[0093] <変形例>

なお、本発明のキャッシュメモリは、上記の実施の形態の構成に限るものではなく、種々の変形が可能である。以下、変形例のいくつかについて説明する。

(1) 定数レジスタ409、コマンドにより設定される構成を示したが、(a)デフォルトとしてインクリメント値またはデクリメント値を保持する、(b)複数回のロード／ストア命令における2つのメモリアクセスアドレスの差分を算出し、インクリメント値またはデクリメント値として保持する、(c)プロセッサにより指定されたアドレス方向(アドレス昇順かアドレス降順か)に応じて上記(b)を算出し保持する構成としてもよい。

(2) プリフェッチ部414において、次のラインのさらに次のライン等複数ラインをプリフェッチするようにしてもよい。同様にタッチ部415も、複数ラインをタッチするようにして

もよい。

(3)プリフェッチ部414は、比較器412で複数回一致した場合にプリフェッチを開始するようにしてもよい。タッチ部415も同様である。

(4)上記実施の形態では、4ウェイ・セット・アソシエイティブのキャッシュメモリを例に説明したが、ウェイ数は、いくつでもよい。また、上記実施の形態では、セット数が16である例を説明したが、セット数はいくつでもよい。

(5)上記実施の形態では、セット・アソシエイティブのキャッシュメモリを例に説明したが、フル・アソシエイティブ方式やダイレクトマップ方式のキャッシュメモリであってもよい。

(6)上記実施の形態では、サブラインのサイズをラインのサイズの1/4としているが、1/2、1/8、1/16等他のサイズでもよい。その場合、各キャッシュエントリーは、サブラインと同数のバリッドフラグおよびダーティフラグをそれぞれ保持すればよい。

(7)上記実施の形態におけるサイズレジスタ403は、サイズを保持する代わりにプリフェッチすべき回数を保持する構成としてもよい。この場合には、予測処理部39は実際にプリフェッチした回数をカウントし、カウント値がプリフェッチすべき回数に達したとき、プリフェッチを停止する構成とすればよい。このようにプリフェッチ回数をカウントすることにより、プロセッサにおけるプログラム進行状態を監視することができる。

[0094] (実施の形態2)

第1の実施の形態では、説明した予測プリフェッチ及び予測タッチでは今後アクセスするであろうラインを推定する場合の構成を説明した。本実施の形態では、既にアクセスしたであろうラインを推定して予測クリーニング(ライトバック)や予測ウィーク化(アクセス順位の最弱化)や予測無効化も行う構成について説明する。

[0095] <キャッシュメモリの構成>

図9は、本発明の実施の形態2におけるキャッシュメモリの構成を示すブロック図である。同図のキャッシュメモリは、図2に示した構成と比較して、ウェイ31a〜31dの代わりにウェイ131a〜131dを備える点と、制御部38の代わりに制御部138を備える点とが異なっている。以下、同じ点は説明を省略して、異なる点を中心に説明する。

[0096] ウェイ131aは、ウェイ31aと比べて、各キャッシュエントリー中に、Cフラグ、Wフラグ

及びUフラグが追加されている点と、ライン単位のバリッドフラグVの代わりにサブライン毎のバリッドフラグV0〜V3を有する点と、ライン単位のダーティフラグDの代わりにサブライン毎のダーティフラグD0〜D3を有する点とが異なる。ウェイ131b〜131dも同様である。

- [0097] 図10に、キャッシュエントリーのビット構成を示す。1つのキャッシュエントリーは、バリッドフラグV0〜V3、21ビットのタグ、128バイトのラインデータ、ウィークフラグW、使用フラグU及びダーティフラグD0〜D3を保持する。
- [0098] タグは21ビットのタグアドレスのコピーである。
ラインデータは、タグアドレスおよびセットインデックスにより特定されるブロック中の128バイトデータのコピーであり、32バイトの4つのサブラインからなる。
- [0099] バリッドフラグV0〜V3は、4つのサブラインに対応し、サブラインが有効か否かを示す。
- [0100] Cフラグ(クリーニングフラグ)は、キャッシングを終了してよいかどうかを示すキャッシング終了属性を示す。C=0は、以降に書き込みがなされる可能性があることを意味する。C=1は、以降に書き込みがなされないことを意味し、ダーティであればクリーニング(ライトバック)によりキャッシングを終了すべきであることを意味する。
- [0101] ウィークフラグWは、プロセッサからのアクセスに関しては、これ以上使用するか否かを示し、キャッシュメモリにおけるリプレース制御に関しては、他のキャッシュエントリーよりも先に追い出してもよい最弱のリプレース対象であることを示す。
- [0102] 使用フラグUは、そのキャッシュエントリーにアクセスがあったか否かを示し、LRU方式におけるキャッシュエントリー間のアクセス順序データの代わりに用いられる。より正確には、使用フラグUの1は、アクセスがあったことを、0はないことを意味する。ただし、1つのセット内の4つウェイの使用フラグが全て1になれば、0にリセットされる。別言すれば、使用フラグUは、アクセスされた時期が古いか新しいか2つの相対的な状態を示す。つまり、使用フラグUが1のキャッシュエントリーは、使用フラグが0のキャッシュエントリーよりも新しくアクセスされたことを意味する。
- [0103] ダーティフラグD0〜D3は、4つのサブラインに対応し、そのサブラインにプロセッサから書き込みがあったか否か、つまりサブライン中にキャッシュされたデータが存在す

るが書き込みによりメモリ中のデータと異なるためメモリに書き戻すことが必要か否かを示す。

[0104] 制御部138は、制御部38と比べて、予測処理部39の代わりに予測処理部139を有する点が異なり、予測によりCフラグを設定する点と、予測によりWフラグを設定する点と、LRU方式におけるアクセス順序情報の代わりに使用フラグUを用いる点とが異なる。

[0105] <使用フラグUの説明>

図11は、制御部138による使用フラグの更新例を示す。同図の上段、中段、下段は、ウェイ0〜3に跨るセットNを構成する4つのキャッシュエントリーを示している。4つのキャッシュエントリー右端の1又は0は、それぞれ使用フラグの値である。この4つの使用フラグUをU0〜U3と記す。

[0106] 同図上段では $(U0 \sim U3) = (1, 0, 1, 0)$ であるので、ウェイ0、2のキャッシュエントリーはアクセスがあったことを、ウェイ1、3のキャッシュエントリーはアクセスがないことを意味する。

[0107] この状態で、メモリアクセスがセットN内のウェイ1のキャッシュエントリーにヒットした場合、同図中段に示すように、 $(U0 \sim U3) = (1, 1, 1, 0)$ に更新される。つまり、実線に示すようにウェイ1の使用フラグU1が0から1に更新される。

[0108] さらに、同図中段の状態で、メモリアクセスがセットN内のウェイ3のキャッシュエントリーにヒットした場合、同図下断に示すように、 $(U0 \sim U3) = (0, 0, 0, 1)$ に更新される。つまり、実線に示すようにウェイ3の使用フラグU1が0から1に更新される。加えて、破線に示すようにウェイ3以外の使用フラグU0〜U2が1から0に更新される。これにより、ウェイ3のキャッシュエントリーが、ウェイ0〜2の各キャッシュエントリーよりも新しくアクセスされたことを意味することになる。

[0109] 制御部138は、キャッシュミス時に $W=1$ のキャッシュエントリーが存在しなければ、使用フラグに基づいてリプレース対象のキャッシュエントリーを決定してリプレースを行う。例えば、制御部138は、図5上段では、ウェイ1とウェイ3の何れかをリプレース対象と決定し、図5中段ではウェイ3をリプレース対象と決定し、図5下段ではウェイ0〜2の何れかをリプレース対象と決定する。

[0110] <ウィークフラグWの説明>

図12(a)ウィークフラグが存在しないと仮定した場合の比較例であり、キャッシュエントリーがリプレースされる様子を示す図である。同図においても、図11と同様にウェイ0〜3に跨るセットNを構成する4つのキャッシュエントリーを示している。、4つのキャッシュエントリー右端の1又は0は、それぞれ使用フラグの値である。また、データEのみアクセス頻度の低いデータを、データA、B、C、Dはアクセス頻度の高いデータとする。

[0111] 同図(a)の第1段目の状態で、プロセッサ1がデータEにアクセスすると、キャッシュミスが発生する。このキャッシュミスにより、例えば、U=0のキャッシュエントリーの中からアクセス頻度の高いデータCのキャッシュエントリーがアクセス頻度の低いデータEにリプレースされ、第2段目の状態となる。

[0112] 第2段目の状態で、プロセッサ1がデータCにアクセスすると、キャッシュミスが発生する。このキャッシュミスにより、U=0のキャッシュエントリーであるアクセス頻度の高いデータDのキャッシュエントリーがアクセス頻度の高いデータCにリプレースされ、第3段目の状態となる。

[0113] 第3段目の状態で、プロセッサ1がデータDにアクセスすると、キャッシュミスが発生する。このキャッシュミスにより、例えば、アクセス頻度の高いデータCのキャッシュエントリーがアクセス頻度の高いデータDにリプレースされ、第3段目の状態となる。

[0114] 同様に、第4段目でも、使用頻度の低いデータEはリプレース対象として選択されないうで、キャッシュメモリーに残っている。

[0115] 第5段目の状態で、使用頻度の低いデータEは最も古い(U=0)であることから、リプレース対象として選択されて、追い出される。

[0116] このように、擬似LRU方式において(通常のLRU方式においても)、アクセス頻度の低いデータEによって、4ウェイの場合は最悪4回のキャッシュミスを誘発する場合がある。

[0117] 図12(b)は、リプレース処理におけるウィークフラグWの役割を示す説明図である。

[0118] 同図(b)の第1段目の状態(同図(a)の第1段目と同じ)で、プロセッサ1がデータEにアクセスすると、キャッシュミスが発生する。このキャッシュミスにより、例えば、U=0

のキャッシュエントリーの中からアクセス頻度の高いデータCのキャッシュエントリーがアクセス頻度の低いデータEにリプレースされる。このとき、プロセッサ1は、データEのキャッシュエントリーにウィークフラグWを1に設定するものとする。これにより、次のキャッシュミス時にデータEのキャッシュエントリーが真っ先に追い出され、第2段目の状態となる。

[0119] 第2段目の状態で、プロセッサ1がデータCにアクセスすると、キャッシュミスが発生する。このキャッシュミスにより、 $W=1$ のキャッシュエントリーであるアクセス頻度の低いデータEのキャッシュエントリーがリプレース対象として選択され、アクセス頻度の高いデータCにリプレースされ、第3段目の状態となる。

[0120] このように、ウィークフラグWを設けることにより、アクセス頻度の低いデータによるキャッシュミスの誘発を低減することができる。

[0121] <予測処理部の構成>

図13は、予測処理部139の構成を示すブロック図である。同図の予測処理部139は、図3に示した予測処理部39と比較して、実行部413の代わりに実行部413aを備える点が異なっている。実行部413aは、実行部413と比べて、C設定部416、W設定部417が追加されている。

[0122] C設定部416は、プロセッサのストア命令の進行状況に応じてストアが完了した直前のラインを推定し、推定したラインのCフラグを1に設定する。

[0123] W設定部417は、プロセッサのストア命令の進行状況に応じてストアが完了した直前のラインを推定し、推定したラインのWフラグを1に設定する。

[0124] <Cフラグ設定処理の説明図>

図14(a)は、C設定部416によるCフラグ設定処理の説明図である。同図(a)において、ラインN、N+1、・・・、N+n、スタートライン、エンドラインは、図6(a)と同様である。

[0125] C設定部416は、プロセッサ1がシーケンシャルにラインN、N+1、・・・に対してデータをストアしていく場合に、例えば、ラインN+1に対してストア命令ST(1)、ST(2)が実行され、ST(2)のメモリアクセスアドレスと予測値とが一致した場合に、ラインNに対するシーケンシャルアクセスによるストアが完了したものと推定し、ラインNのCフ

ラグを1に設定する。Cフラグが1に設定されたラインNは、キャッシュミスの発生を待たずにクリーニング処理によりライトバックされる。

[0126] 同様、ラインN+2へのストア命令実行中に予測値を一致した場合、ラインN+1に対するシーケンシャルアクセスによるストアが完了したものと推定し、ラインN+1のCフラグを1に設定する。

[0127] このように、C設定部416は、Cフラグの設定をスタートラインからエンドラインまでの範囲内で実行する。

[0128] <Wフラグ設定処理の説明図>

図14(b)は、W設定部417によるWフラグ設定処理の説明図である。同図(a)と比べて、Cフラグの代わりにWフラグを設定する点のみが異なり、これ以外は同様であるので説明を省略する。

[0129] Wフラグが1に設定されたラインは、キャッシュミスの発生時に真っ先にリプレース対象として選択され、キャッシュメモリから追い出される。

[0130] <Cフラグ設定処理フロー>

図15は、C設定部416におけるCフラグ設定処理の一例を示すフローチャートである。

同図において、C設定部416は、コマンドレジスタ401にCフラグ設定コマンドが保持されていて(S41b)、プロセッサがストア命令を実行したとき(S42b)、当該メモリアクセスアドレスと定数とを加算した値を初期値として予測値レジスタ408に設定する(S43)。この最初のストア命令のメモリアクセスアドレスと予測値レジスタの予測値とは当然一致しない。予測値レジスタ408はクリアされているかランダムな値を保持しているからである。

[0131] さらに、プロセッサがストア命令を実行し(S44b)、かつメモリアクセスアドレスと予測値レジスタの予測値とが一致する場合には(S45)、C設定部416は、直前のラインのラインアドレスを算出し(S46b)、算出したラインアドレスがスタートラインからエンドラインまでのアドレス範囲に属していて(S47)、直前のラインがキャッシュメモリにエンタリーされていれば(S48b)、直前のラインのCフラグを1に設定する(S49b)。

[0132] さらに、C設定部416は、直前のラインがエンドラインであれば(S50)Cフラグ設定

処理を終了する。

- [0133] このように、C設定部416は、プロセッサのストア命令の進行状況に応じて、ストアが完了した直前のラインを推定し、推定したラインのCフラグを1に設定するので、プロセッサ1では、キャッシュメモリのラインサイズやライン境界を認識する必要がなく、つまり、クリーニングしてもよいラインをラインアドレスとしてプロセッサ1にて管理する必要がない。キャッシュメモリにおいて推定に基づいて効率よくクリーニングを行うことができる。

- [0134] <Wフラグ設定処理フロー>

図16は、W設定部417におけるWフラグ設定処理の一例を示すフローチャートである。同図のフローは、図15のCフラグ設定処理と比較して、Cフラグの代わりにWフラグを設定する点のみが異なり、これ以外は同様であるので説明を省略する。

- [0135] このように、W設定部417は、プロセッサのストア命令の進行状況に応じて、ストアが完了した直前のラインを推定し、推定したラインのWフラグを1に設定するので、キャッシュメモリにおいて推定に基づいてストアが完了した直前のラインを効率よくリプレース対象とすることができる。その際、プロセッサ1では、キャッシュメモリのラインサイズやライン境界を認識する必要がなく、つまり、リプレースしてもよいラインをラインアドレスとしてプロセッサ1にて管理する必要がない。

- [0136] <クリーニング処理>

図17は、制御部138におけるクリーニング処理の一例を示すフローチャートである。

同図のように、制御部138は、ループ1の処理(S900〜S913)において、セットインデックス(SI)0〜15を順に指定する(S901)ことにより、16個のすべてのセットに対してループ2の処理を行う。ループ2の処理(S900〜S913)において、制御部138は、セット内のウェイ0〜3のCフラグを読み出す(S903)ことにより、C=1のキャッシュエントリを探索する(S904)。ループ3の処理(S905〜910)において、制御部138は、C=1のキャッシュエントリに対して、サブライン単位のダーティフラグを読み出し(S906)、ダーティであれば(S907)、そのサブランのデータをメモリ2に書き戻し(S908)、当該ダーティフラグを0にリセットする(S909)。このサブラインデータの書

き戻しにおいて、制御部138は、ループ4の処理(S920〜S923)のように、空きサイクルにおいて(S920)、1ワードずつ書き戻す(S922)

このように、制御部138は、全てのキャッシュエントリーのCフラグを順にチェックして、C=1のキャッシュエントリーを探索し、ダーティであればキャッシュメモリ3からメモリ2に書き戻す。

[0137] このように、クリーニング処理では、これ以上書き込みされないことを示すCフラグを有するキャッシュエントリーを、キャッシュミスが発生する前にライトバックするので、キャッシュミス時にはロードペナルティが発生するだけでライトバックペナルティの発生を低減することができる。これによりキャッシュメモリの効率を向上させ、アクセス速度を向上させることができる。

[0138] <Uフラグ更新処理>

図18は、制御部138におけるUフラグ更新処理を示すフローチャートである。同図では、バリッドフラグが0(無効)であるキャッシュエントリーの使用フラグUは0に初期化されているものとする。

[0139] 同図において、制御部138は、キャッシュヒットしたとき(ステップS61)、セットインデックスにより選択されたセットにおけるヒットしたウェイの使用フラグUを1にセットし(ステップS62)、そのセット内の他のウェイの使用フラグUを読み出し(ステップS63)、読み出した使用フラグUが全て1であるか否かを判定し(ステップS64)、全て1でなければ終了し、全て1であれば他のウェイの全ての使用フラグUを0にリセットする(ステップS65)。

[0140] このようにして制御部138は、図11、図12(a) (b)に示した更新例のように、使用フラグUを更新する。

[0141] <リプレース処理>

図19は、制御部138におけるリプレース処理を示すフローチャートである。同図において制御部138は、メモリアクセスがミスしたとき(ステップS91)、セットインデックスにより選択されたセットにおける、4つウェイの使用フラグU及びウィークフラグWを読み出し(ステップS92)、W=1のウェイが存在するか否かを判定する(ステップS93)。W=1のウェイが存在しないと判定された場合、U=0のウェイを1つ選択する(ステ

ップS94)。このとき、使用フラグUが0になっているウェイが複数存在する場合は、制御部138はランダムに1つを選択する。また、 $W=1$ のウェイが存在すると判定された場合、Uフラグの値に関わらず $W=1$ のウェイを1つ選択する(ステップS95)。このとき、ウィークフラグWが1になっているウェイが複数存在する場合は、制御部138はランダムに1つを選択する。

[0142] さらに、制御部138は、当該セットにおける選択されたウェイのキャッシュエントリーを対象にリプレースし(ステップS96)、リプレース後に当該キャッシュエントリーの使用フラグUを1に、ウィークフラグWを0初期化する(ステップS97)。なお、このときバリッドフラグV、ダーティフラグDは、それぞれ1、0に初期化される。

[0143] このように、 $W=1$ のウェイが存在しない場合、リプレース対象は、使用フラグUが0のキャッシュエントリーの中から1つ選択される。

[0144] また、 $W=1$ のウェイが存在する場合、リプレース対象は、使用フラグUが0であると1であることを問わず、 $W=1$ のウェイのキャッシュエントリーから1つ選択される。これにより図14(a)(b)に示したように、アクセス頻度の低いデータがキャッシュメモリに残ることによるキャッシュミスの誘発を低減することができる。

[0145] 以上説明してきたように、本実施の形態におけるキャッシュメモリによれば、プロセッサ1によるストア命令の進行状態から、ストアが完了してラインを推定し、推定したラインに対してCフラグ又はWフラグを設定するので、ストアが完了した直前のラインを効率よくリプレース対象を指定することができる。その際、プロセッサ1によってキャッシュメモリのライン境界やラインサイズを管理する必要がなく、キャッシュ管理のための負荷を小さくすることができる。

[0146] また、ウィークフラグ $W=1$ でかつダーティフラグ $=1$ のラインを、プロセッサからこれ以上書き込みがなされないラインとして、クリーニングすることにより、キャッシュミスのライトバックペナルティを低減することができる。

[0147] また、これ以上使用されないキャッシュエントリーに $W=1$ が設定され、 $W=1$ のキャッシュエントリーが真っ先にリプレース対象として選択されるので、アクセス頻度の低いデータがキャッシュメモリに残ることによるキャッシュミスの誘発を低減することができる。

[0148] また、従来のLRU方式におけるアクセス順序を示すデータの代わりに1ビットの使用フラグを用いる擬似LRU方式を採用することにより、アクセス順序データとして1ビットのフラグでよいので、アクセス順序データのデータ量が少ないこと及び更新が簡単であることからハードウェア規模を小さくすることができる。

[0149] <変形例>

(1) 図4(a)(b)(c)に示した各命令は、コンパイラによりプログラム中に挿入してもよい。その際、コンパイラは、例えば配列データの書き込みや、圧縮動画データをデコードする際のブロックデータの書き込み等、これ以上書き込みをしないプログラム位置に、上記各命令を挿入するようにすればよい。

(2) 上記クリーニング処理において無効化処理を行う構成としてもよい。すなわち、図17に示したフローチャートにおいて、S907にてダーティでないと判定された場合、さらに、当該キャッシュエントリーを無効化(Vフラグをリセット)するステップを追加してもよい。さらに、クリーニング処理におけるライトバックの後に無効化するステップを追加してもよい。

産業上の利用可能性

[0150] 本発明は、メモリアクセスを高速化するためのキャッシュメモリに適しており、例えば、オンチップキャッシュメモリ、オフチップキャッシュメモリ、データキャッシュメモリ、命令キャッシュメモリ等に適している。

請求の範囲

- [1] プロセッサの状態に関する条件を生成する条件生成手段と、
現在のプロセッサの状態が前記条件を満たすかどうかを判定する判定手段と、
操作対象となるアドレスを生成するアドレス生成手段と、
前記判定手段が条件を満たすと判定したときに前記アドレス生成手段によって生成されたアドレスを用いてキャッシュを操作する操作手段と
を備えることを特徴とするキャッシュメモリシステム。
- [2] 前記条件生成手段は、前記判定手段が条件を満たすと判定した場合に新たな条件を生成する
ことを特徴とする請求項1記載のキャッシュメモリシステム。
- [3] 前記条件生成手段は、プロセッサ内の特定レジスタの値に関する条件を生成する
ことを特徴とする請求項2記載のキャッシュメモリシステム。
- [4] 前記特定レジスタはプログラムカウンタである
ことを特徴とする請求項3記載のキャッシュメモリシステム。
- [5] 前記条件生成手段は、特定のアドレス範囲内へのメモリアクセスおよび特定のアドレス範囲外へのメモリアクセスの何れかを前記条件として生成する
ことを特徴とする請求項2記載のキャッシュメモリシステム。
- [6] 前記条件生成手段は、プロセッサが特定命令を実行することを前記条件として生成する
ことを特徴とする請求項1記載のキャッシュメモリシステム。
- [7] 前記条件生成手段は、現在の条件に特定の演算を施すことによって前記新たな条件を生成する
ことを特徴とする請求項2記載のキャッシュメモリシステム。
- [8] 前記条件生成手段はメモリアクセスアドレスを条件として生成し、
前記判定手段が条件を満たすと判定した場合に現在の条件に定数を加算することによって前記新たな条件を生成する
ことを特徴とする請求項7記載のキャッシュメモリシステム。
- [9] 前記定数は、プロセッサにより実行されるポストインクリメント付きロード／ストア命令

におけるインクリメント値またはデクリメント値、およびプロセッサにより実行される2回のロード／ストア命令におけるアドレスの差分値の何れかである

ことを特徴とする請求項8記載のキャッシュメモリシステム。

- [10] 前記条件生成手段は複数の条件を生成し、
前記判定手段は、複数の条件のすべてを満たすかどうかを判定することを特徴とする請求項1記載のキャッシュメモリシステム。

- [11] 前記条件生成手段は複数の条件を生成し、
前記判定手段は、複数の条件の何れかを満たすかどうかを判定することを特徴とする請求項1記載のキャッシュメモリシステム。

- [12] 前記操作手段は、
前記判定手段が条件を満たすと判定したときに、前記アドレス生成手段により生成されたアドレスに対応するデータがキャッシュに格納されているかどうかを判定するデータ判定手段と、
格納されていないと判定された場合に、キャッシュメモリ中のラインを選択する選択手段と、
前記選択されたラインが有効でダーティならライトバックを行うライトバック手段と、
前記アドレスに対応するデータをメモリからライトバック後の選択されたラインへ転送する転送手段と、
前記アドレスをタグとして前記選択されたラインへ登録する登録手段と
を備えることを特徴とする請求項1から3の何れかに記載のキャッシュメモリシステム。

- [13] 前記操作手段は、
前記判定手段が条件を満たすと判定したときに、前記アドレス生成手段により生成されたアドレスに対応するデータがキャッシュに格納されているかどうかを判定するデータ判定手段と、
格納されていないと判定された場合に、キャッシュメモリ中のラインを選択する選択手段と、
選択されたラインが有効でダーティであれば、ライトバックを行うライトバック手段と、

メモリから選択されたラインヘデータを転送することなく、前記生成したアドレスをタグとして選択されたラインへ登録する登録手段と

を備えることを特徴とする請求項1から3の何れかに記載のキャッシュメモリシステム。

[14] 前記操作手段は、

前記判定手段が条件を満たすと判定したときに、前記アドレス生成手段により生成されたアドレスに対応するデータがキャッシュに格納されているかどうかを判定するデータ判定手段と、

格納されていると判定された場合に、キャッシュメモリ中の格納先のラインを選択する選択手段と、

選択されたラインが有効でかつダーディであればライトバックを行うライトバック手段と、

を備えることを特徴とする請求項1から3の何れかに記載のキャッシュメモリシステム。

[15] 前記操作手段は、

前記判定手段が条件を満たすと判定したときに、前記アドレス生成手段により生成されたアドレスに対応するデータがキャッシュに格納されているかどうかを判定するデータ判定手段と、

格納されていると判定された場合に、キャッシュメモリ中の格納先のラインを選択する選択手段と、

選択されたラインを無効化する無効化手段と

を備えることを特徴とする請求項1から3の何れかに記載のキャッシュメモリシステム。

[16] 前記操作手段は、

前記判定手段が条件を満たすと判定したときに、前記アドレス生成手段により生成されたアドレスに対応するデータがキャッシュに格納されているかどうかを判定するデータ判定手段と、

格納されていると判定された場合に、キャッシュメモリ中の格納先のラインを選択す

る選択手段と、

ラインのアクセス順序を示す順序情報に対して、選択されたラインのアクセス順序を変更する変更手段と、

を備えることを特徴とする請求項1から3の何れかに記載のキャッシュメモリシステム。

[17] 前記条件生成手段により前記条件としてメモリアドレスを生成し、

前記操作手段は、さらに、

前記条件生成手段により生成されたメモリアドレスがラインの途中を指す場合に、当該ラインの先頭、次のラインの先頭および前のラインの先頭の何れかを指すように調整することによりアドレスを生成する調整手段を備える

ことを特徴とする請求項12から16の何れかに記載ののキャッシュシステム。

[18] キャッシュメモリの制御方法であって、

プロセッサの状態に関する条件を生成する条件生成ステップと、

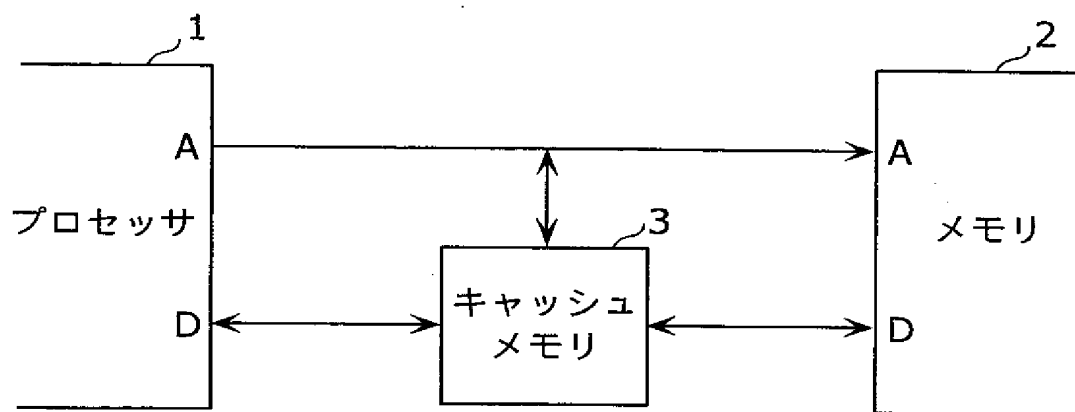
現在のプロセッサの状態が前記条件を満たすかどうかを判定する判定ステップと、

操作対象となるアドレスを生成するアドレス生成ステップと、

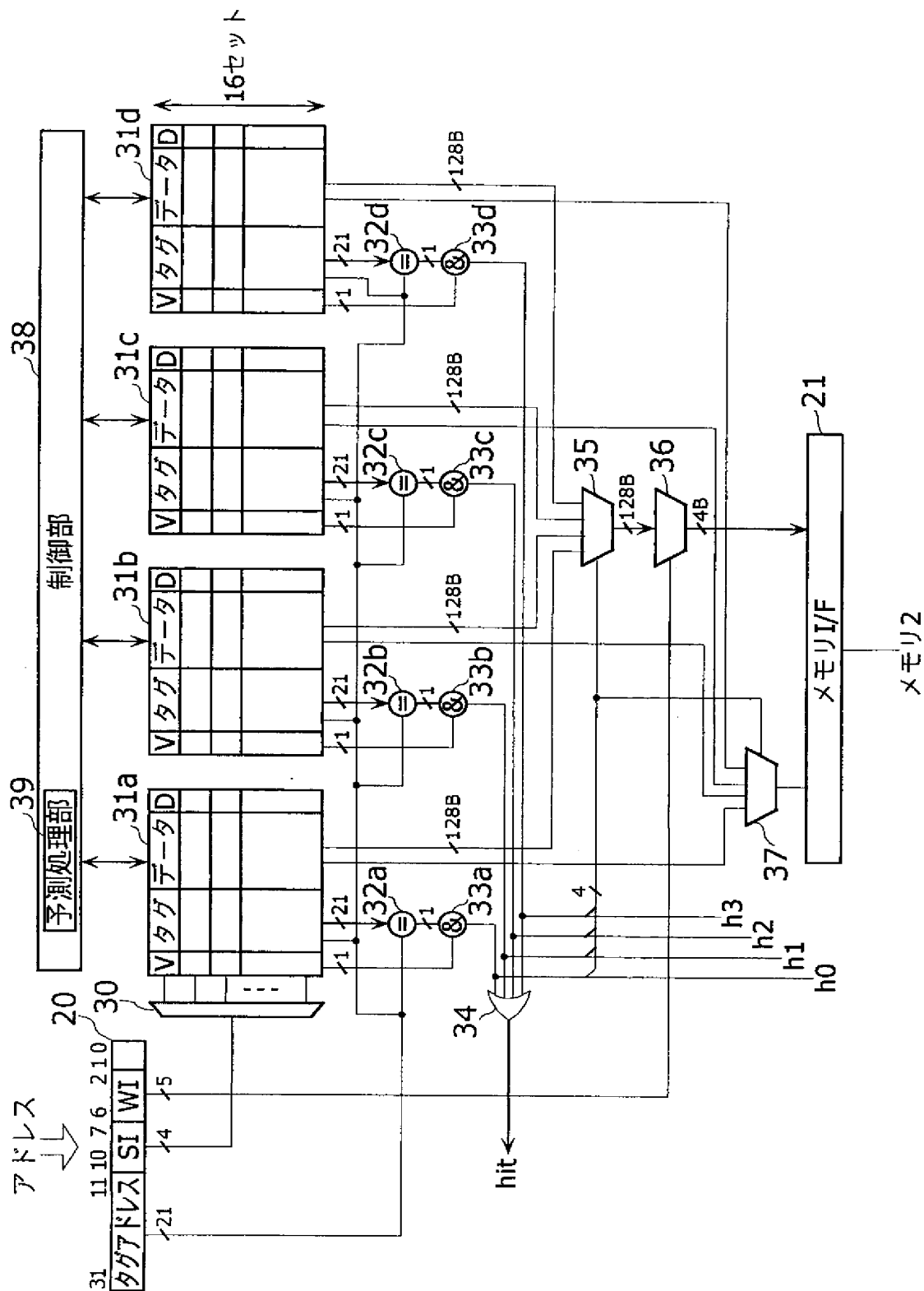
前記判定ステップにおいて条件を満たすと判定したときに前記アドレス生成ステップにおいて生成されたアドレスを用いてキャッシュを操作する操作ステップと

を有することを特徴とする制御方法。

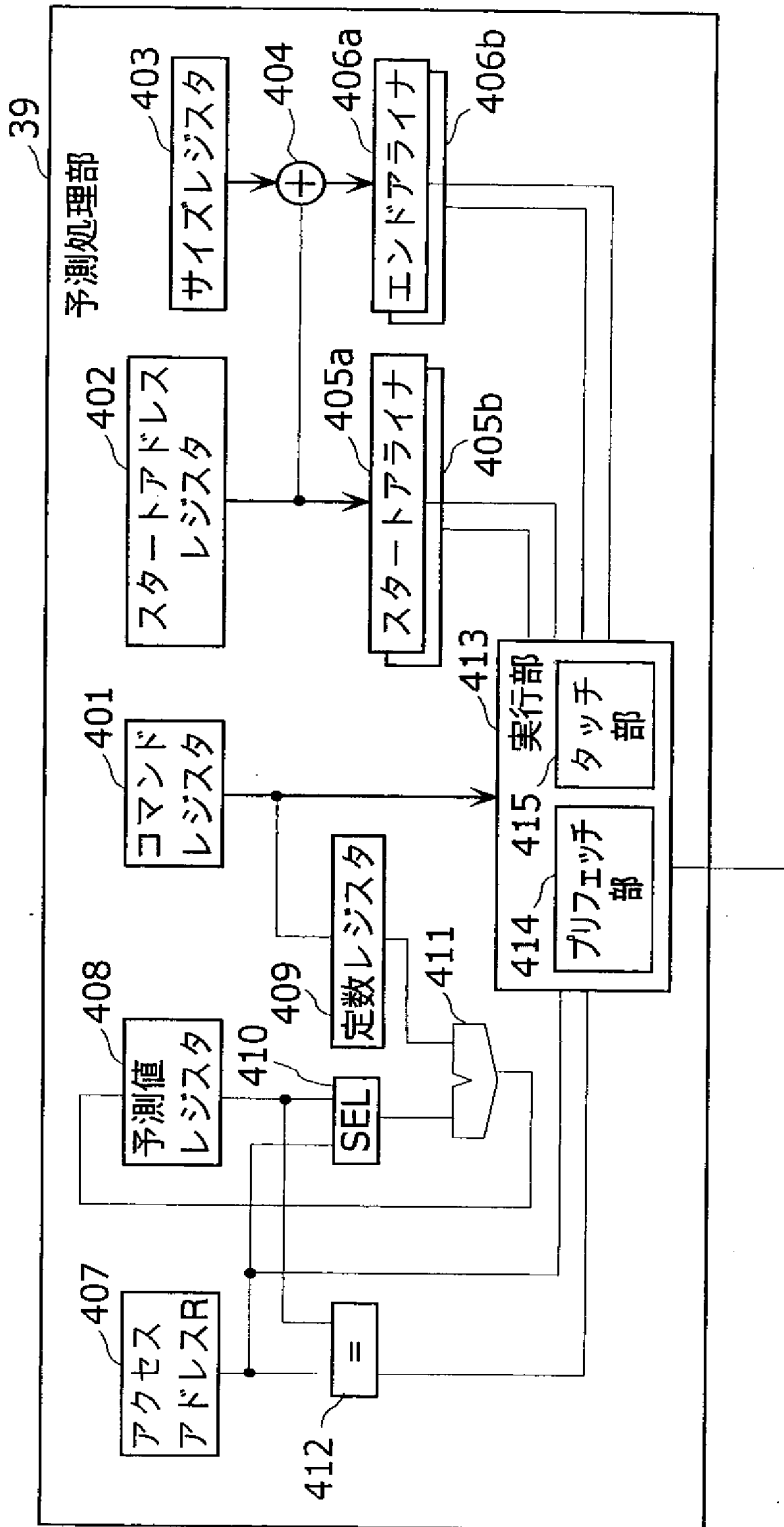
[図1]



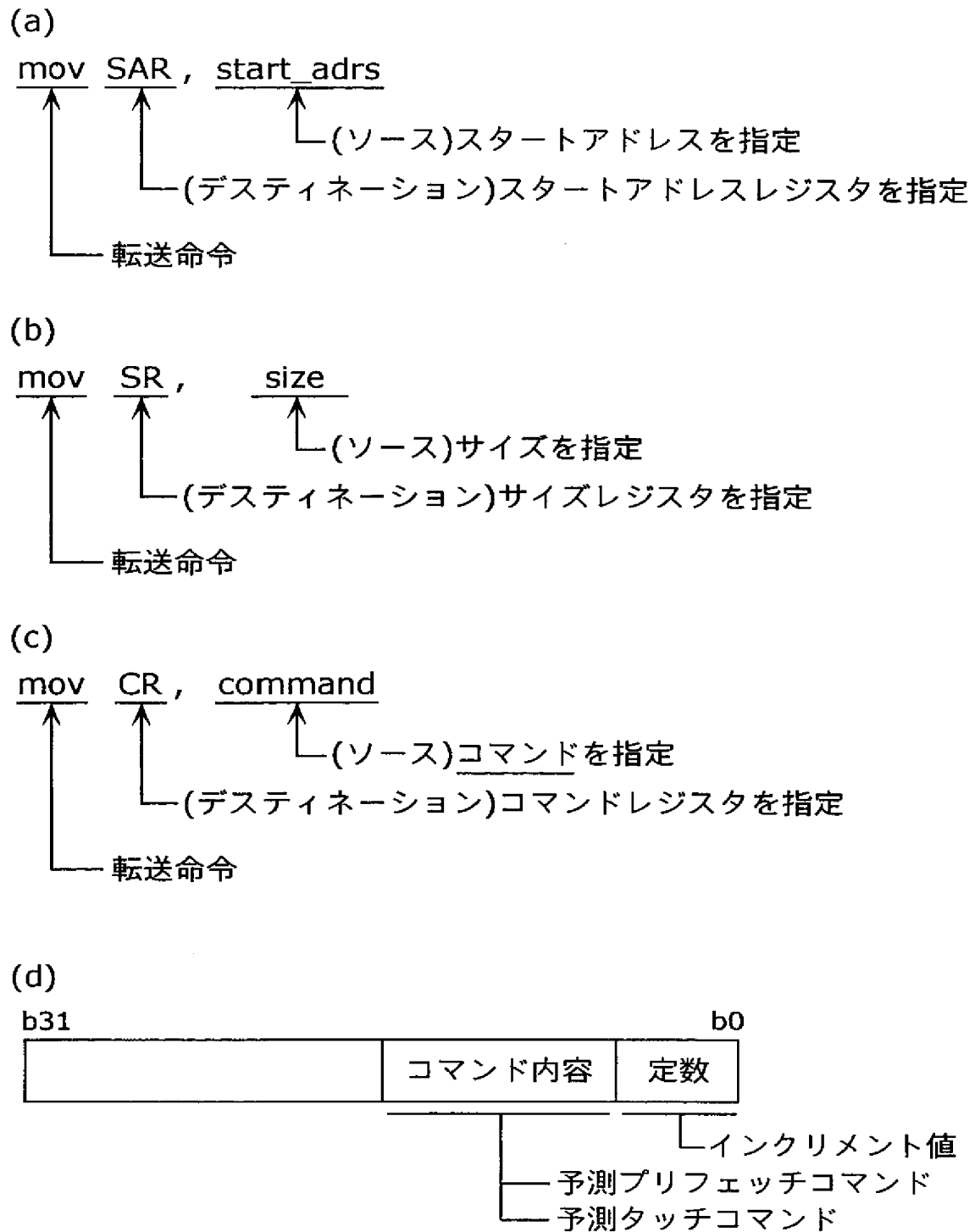
[図2]



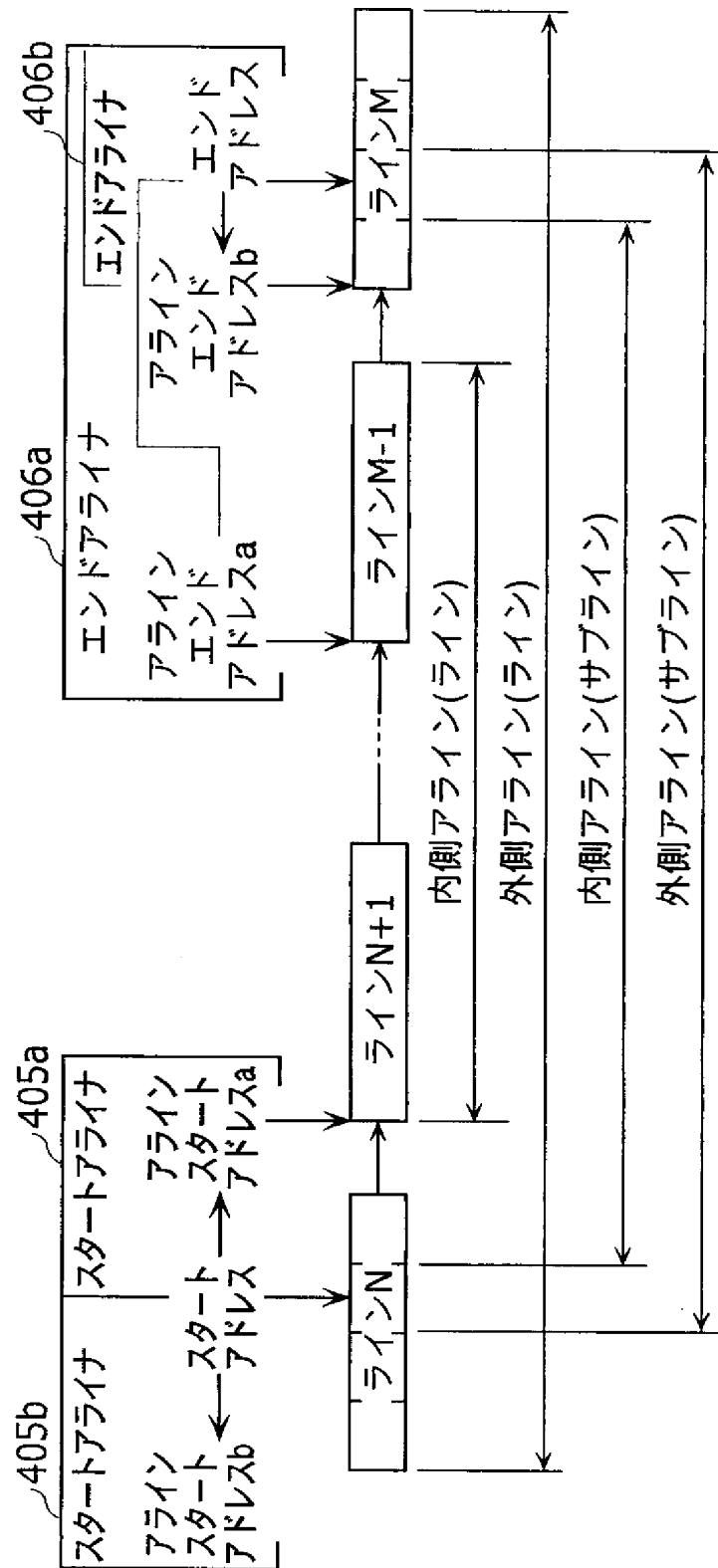
[図3]



[図4]

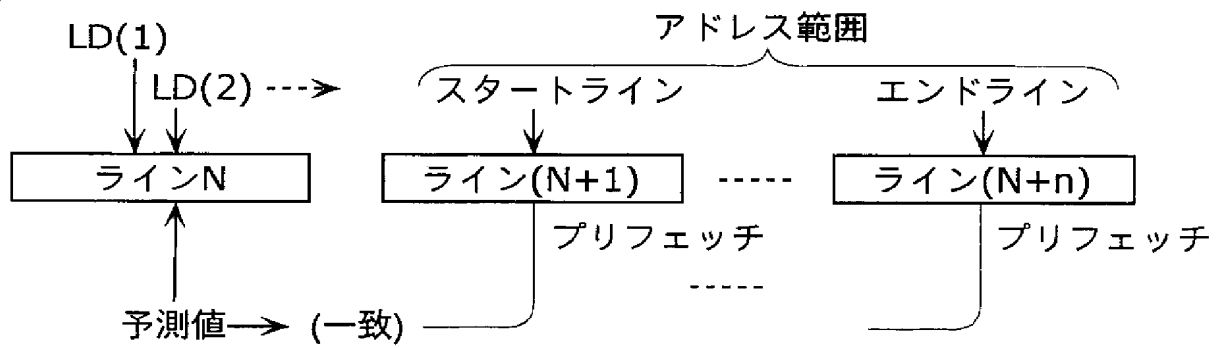


[図5]

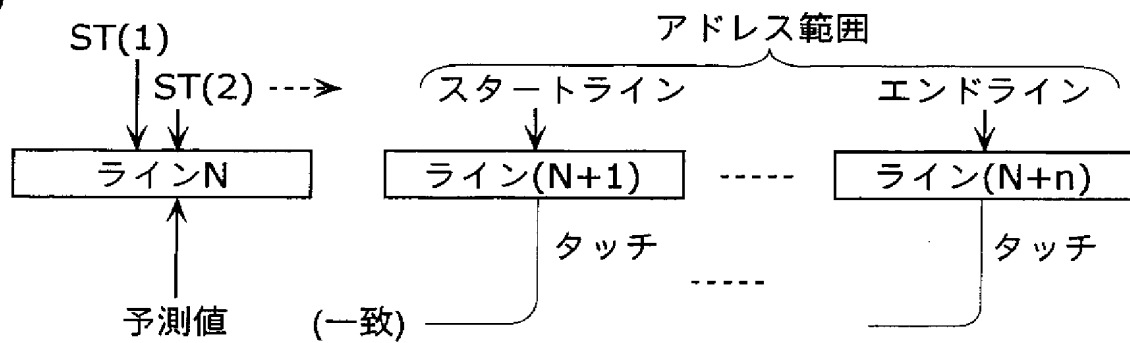


[図6]

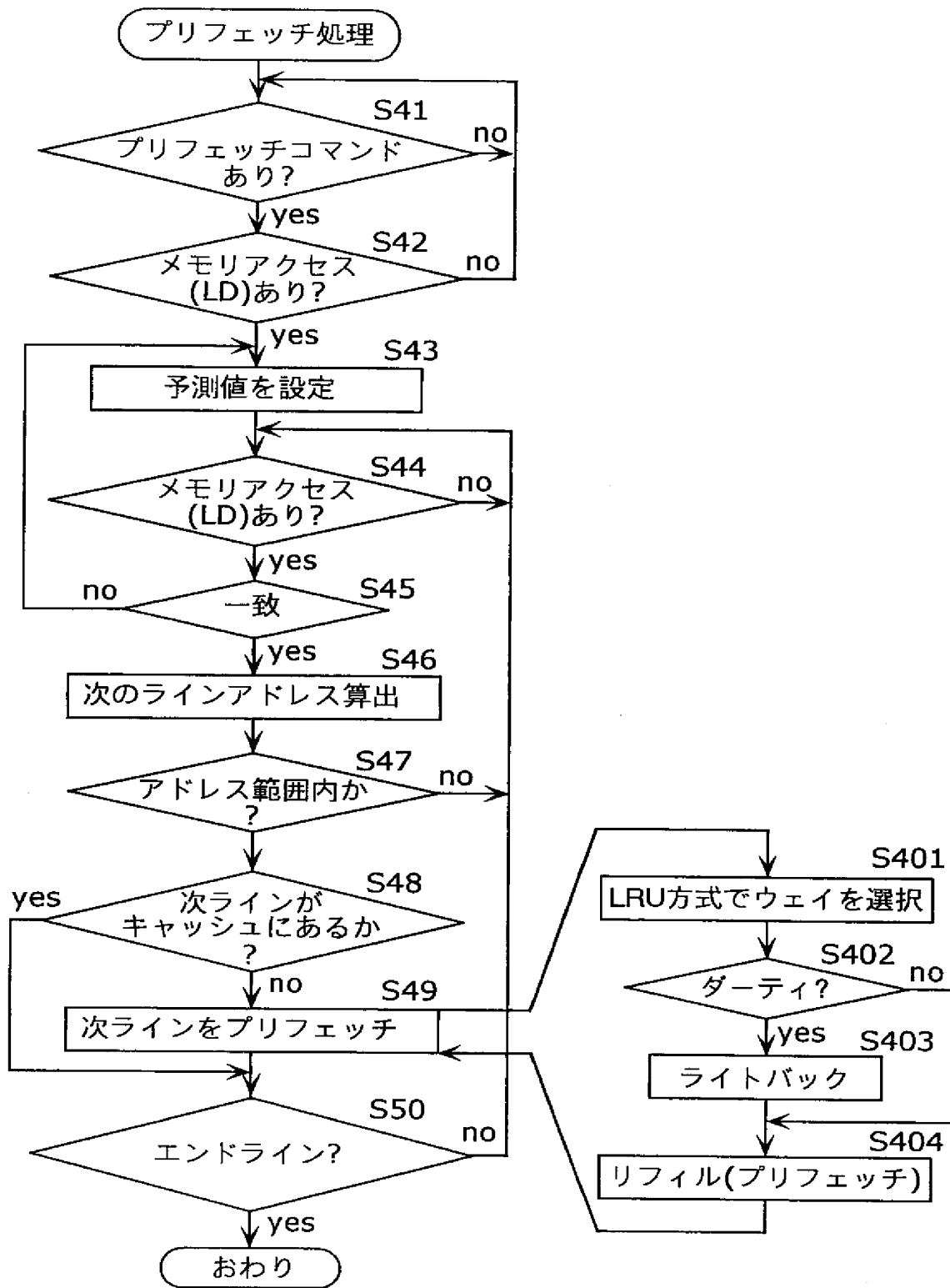
(a)



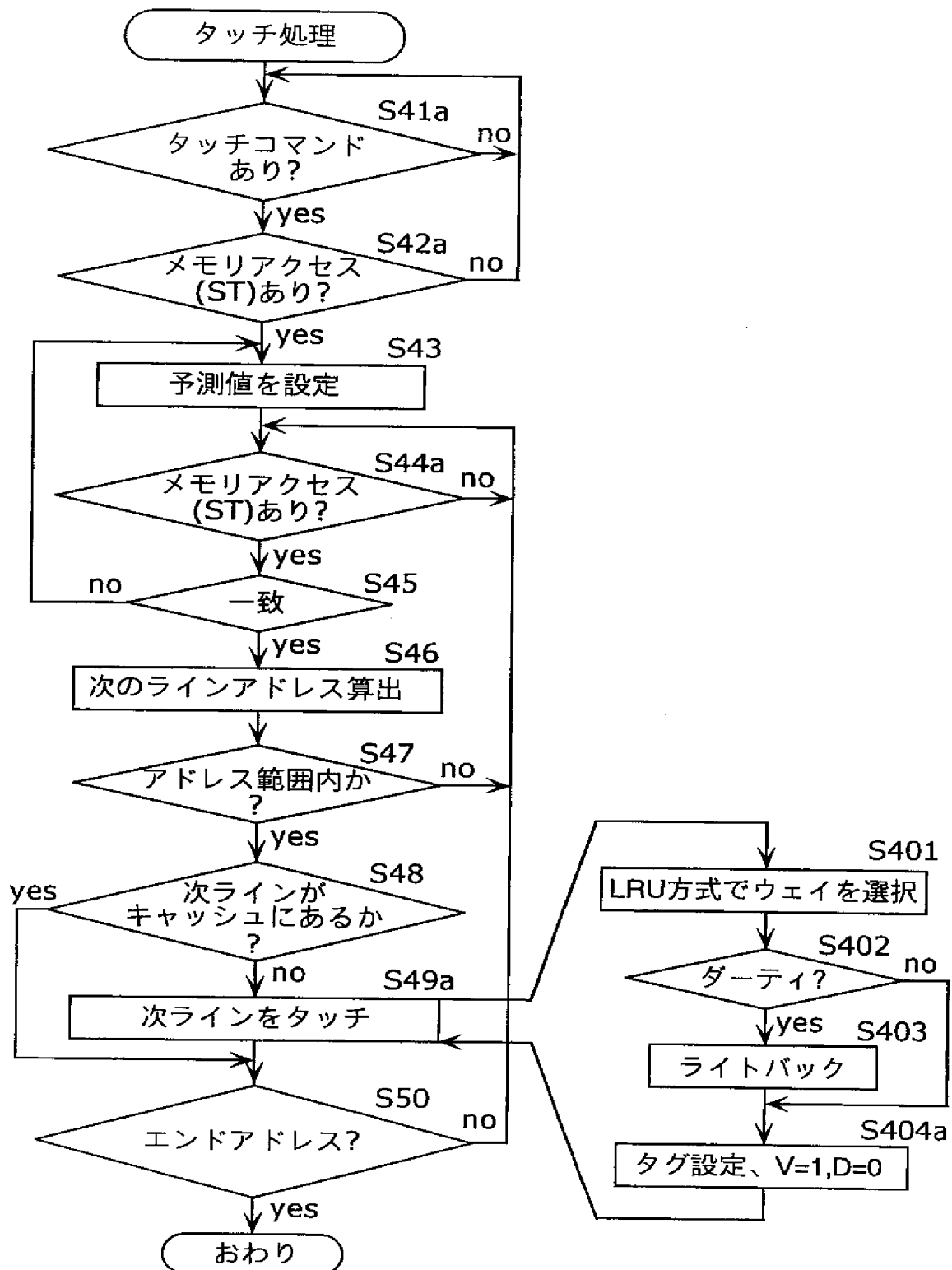
(b)



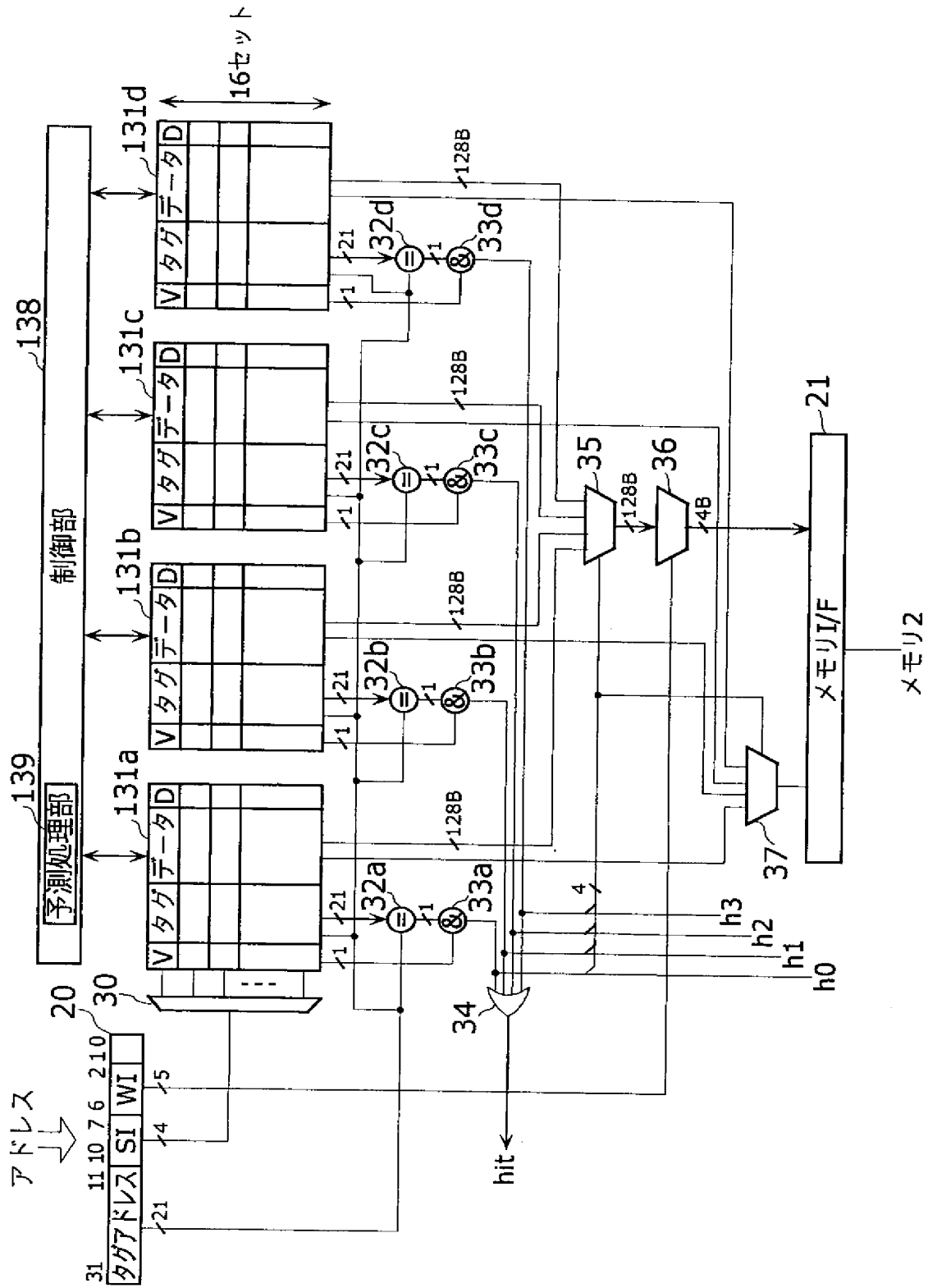
[図7]



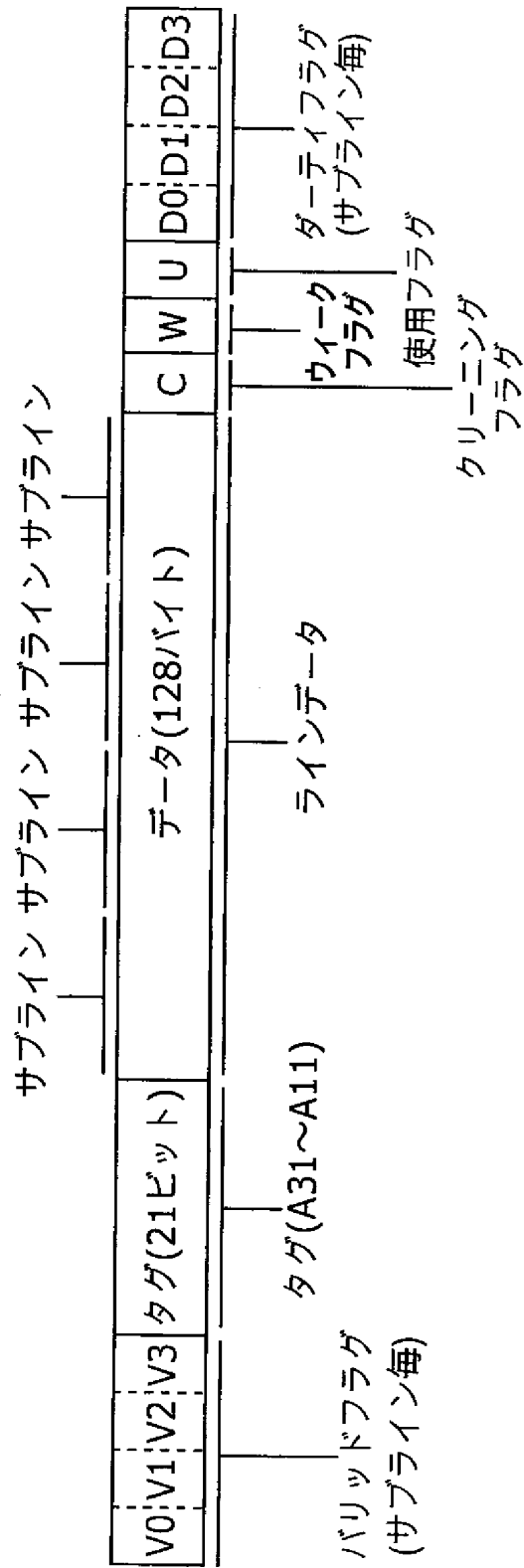
[図8]



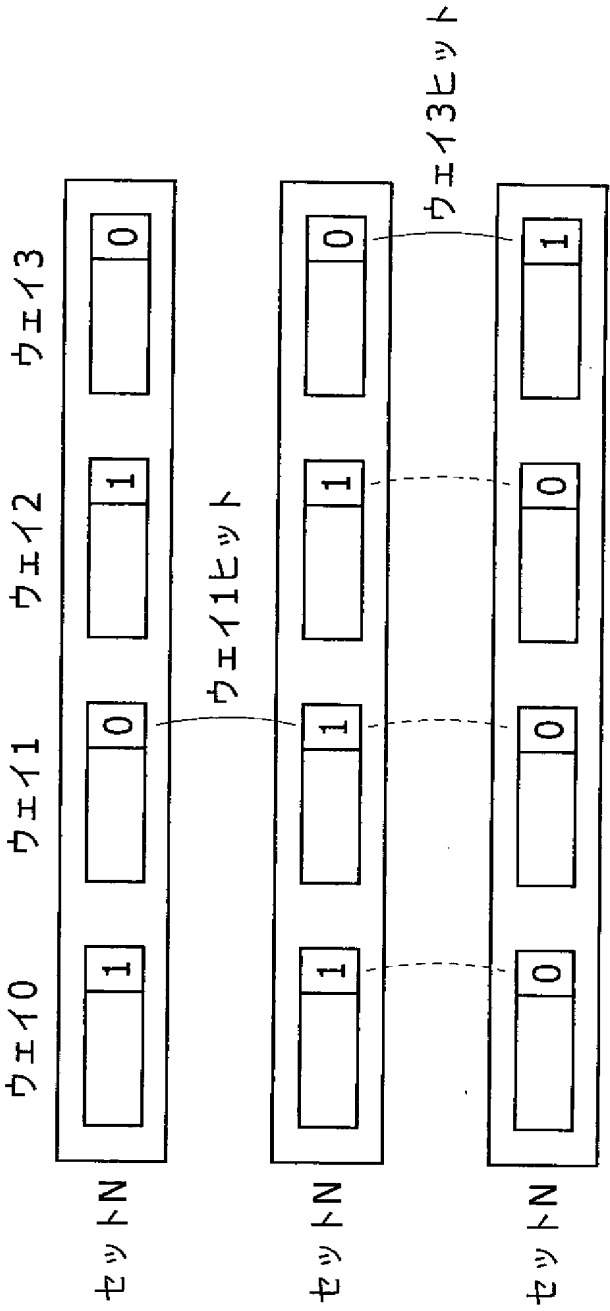
[図9]



[図10]

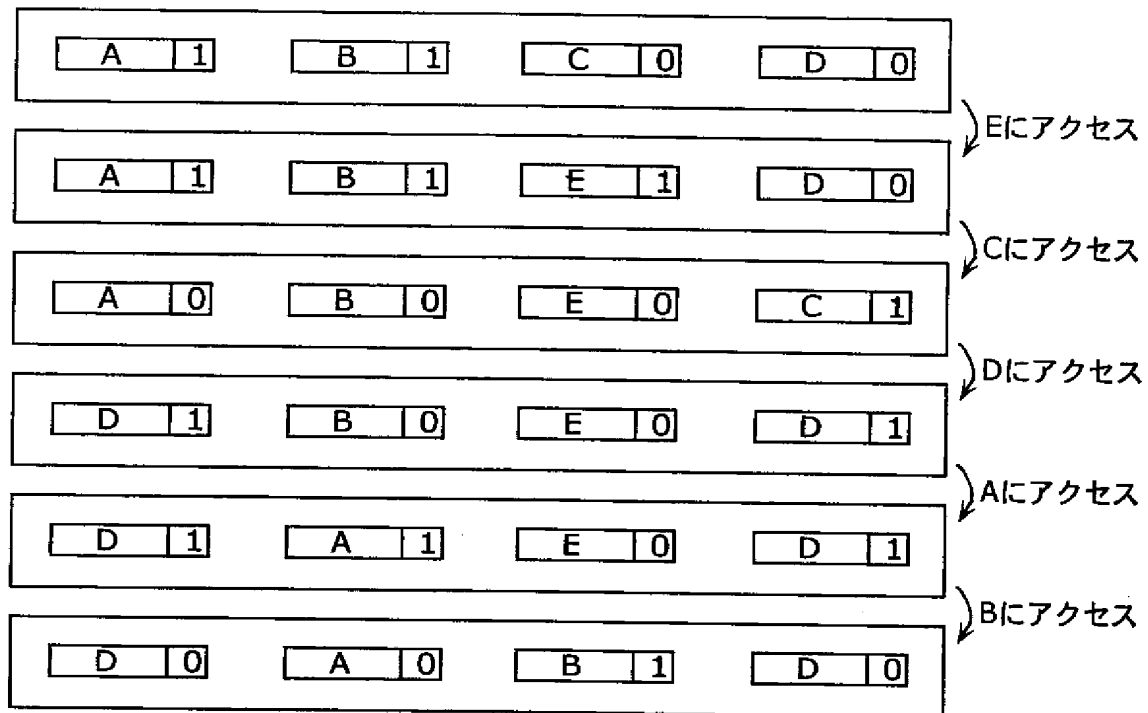


[図11]

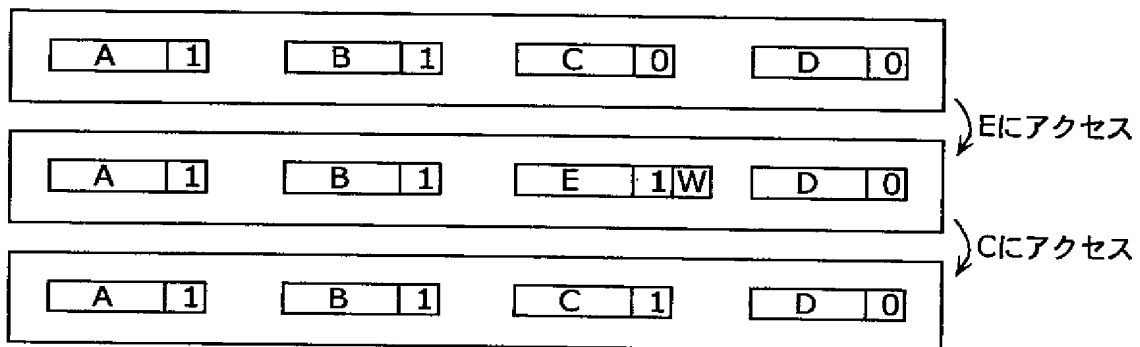


[図12]

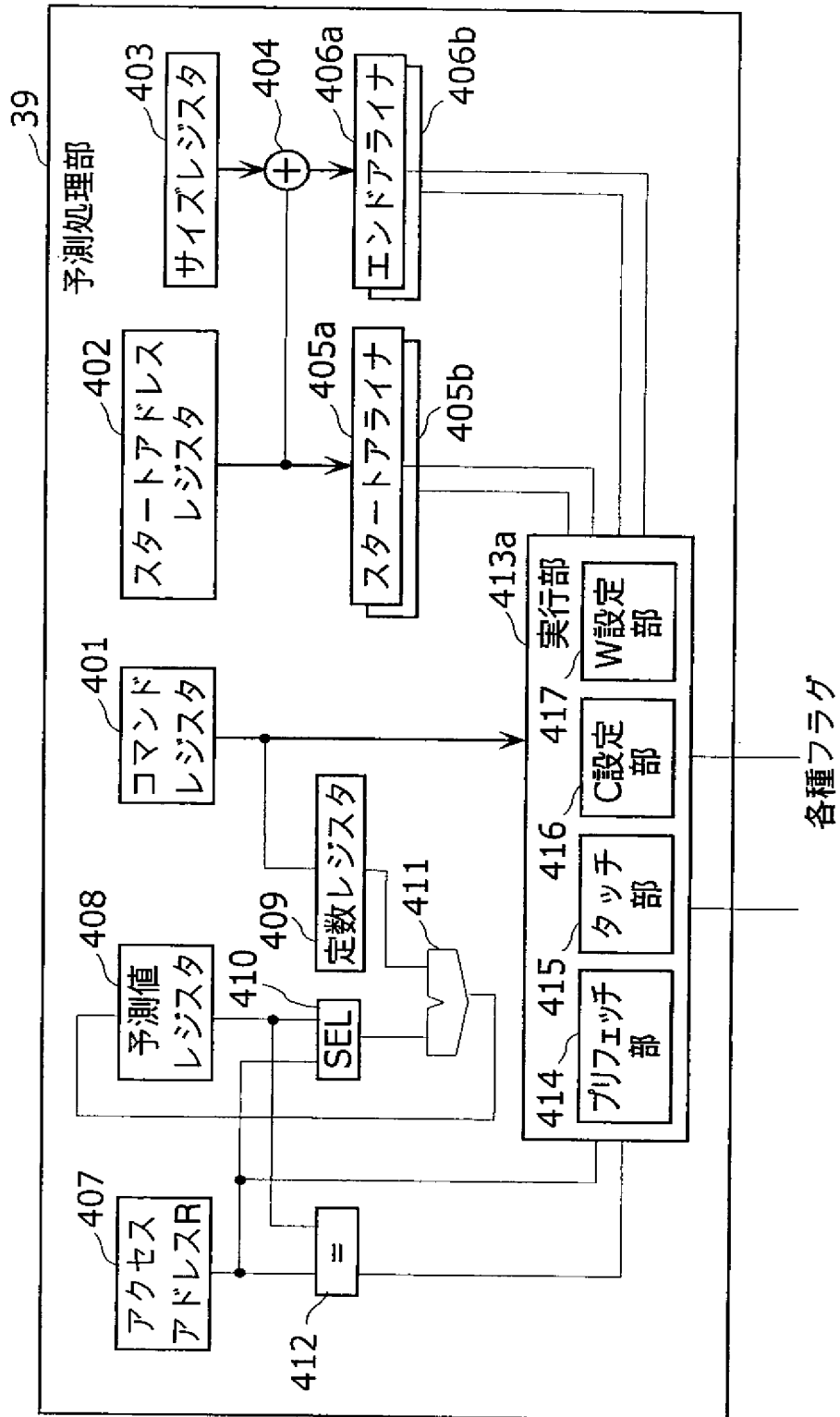
(a)



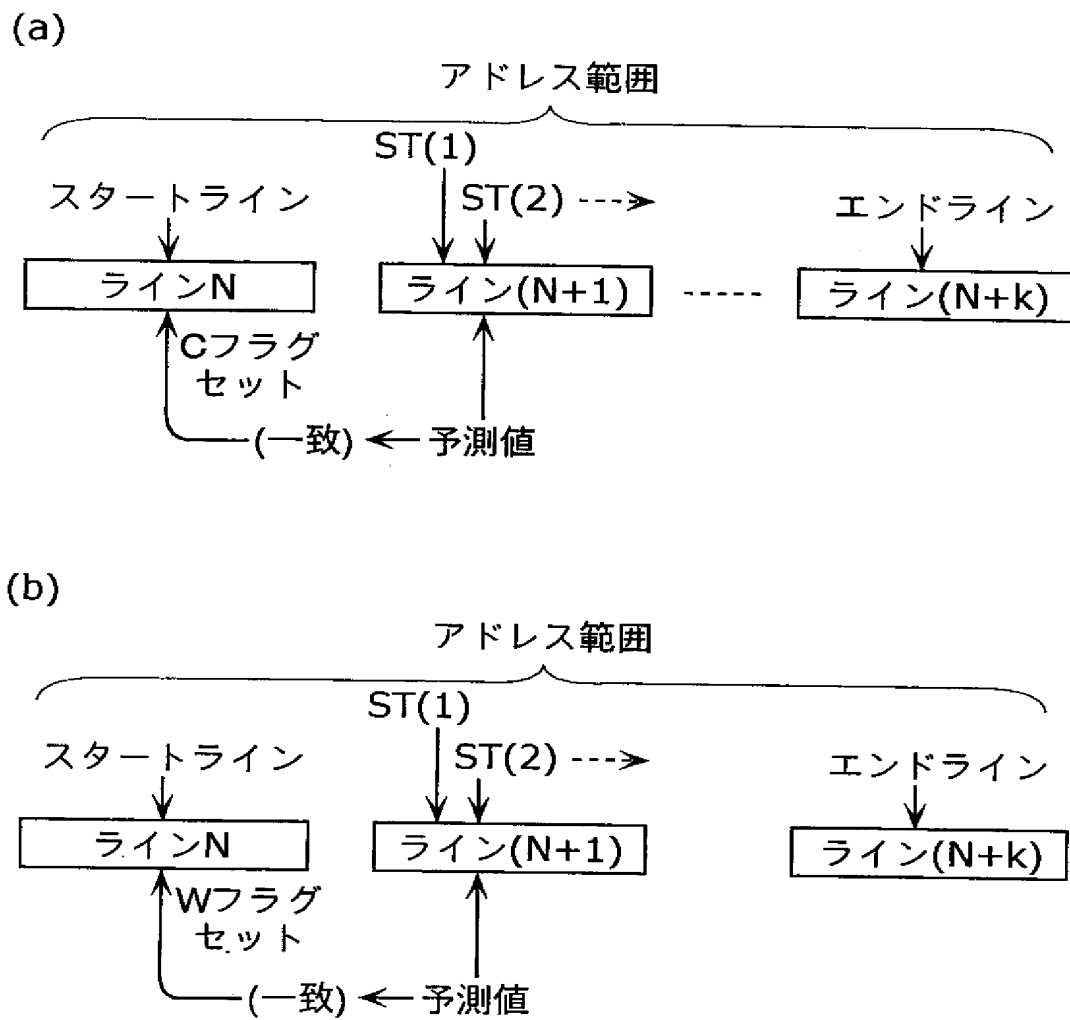
(b)



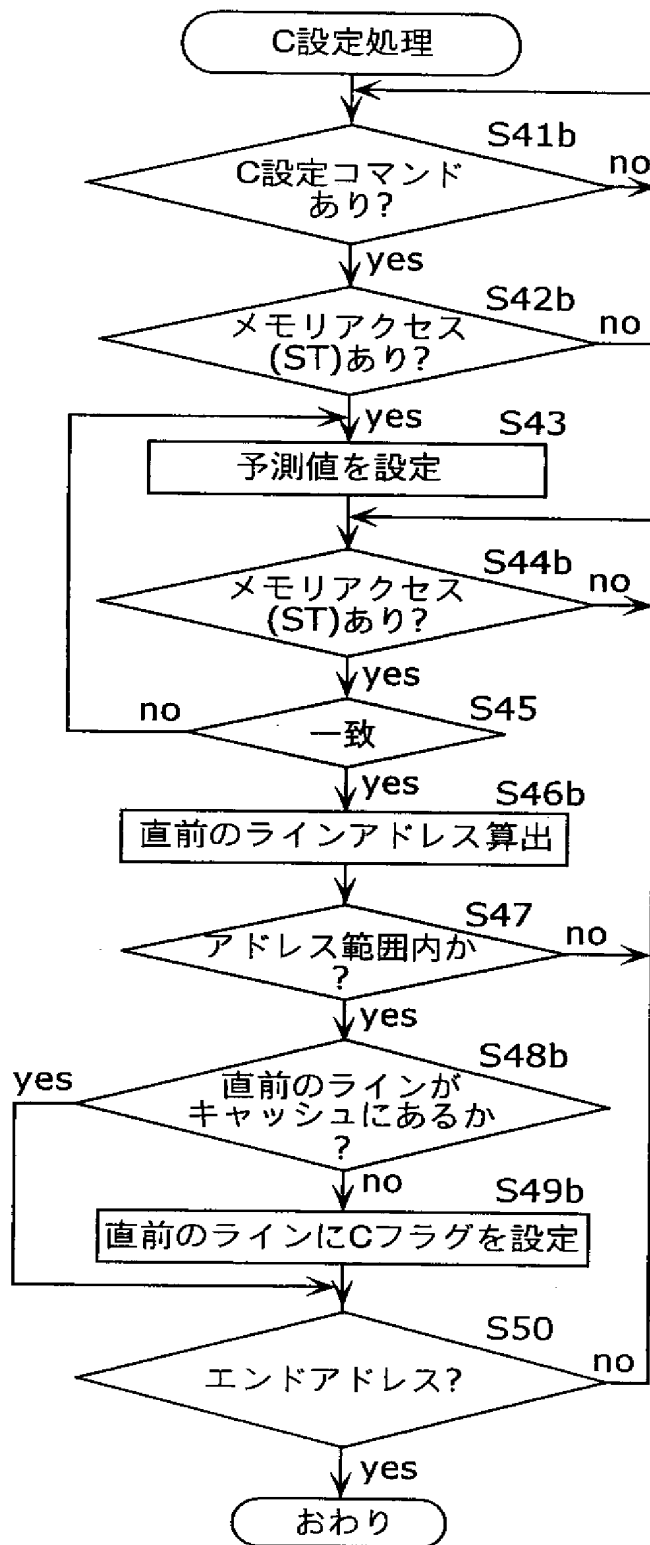
[図13]



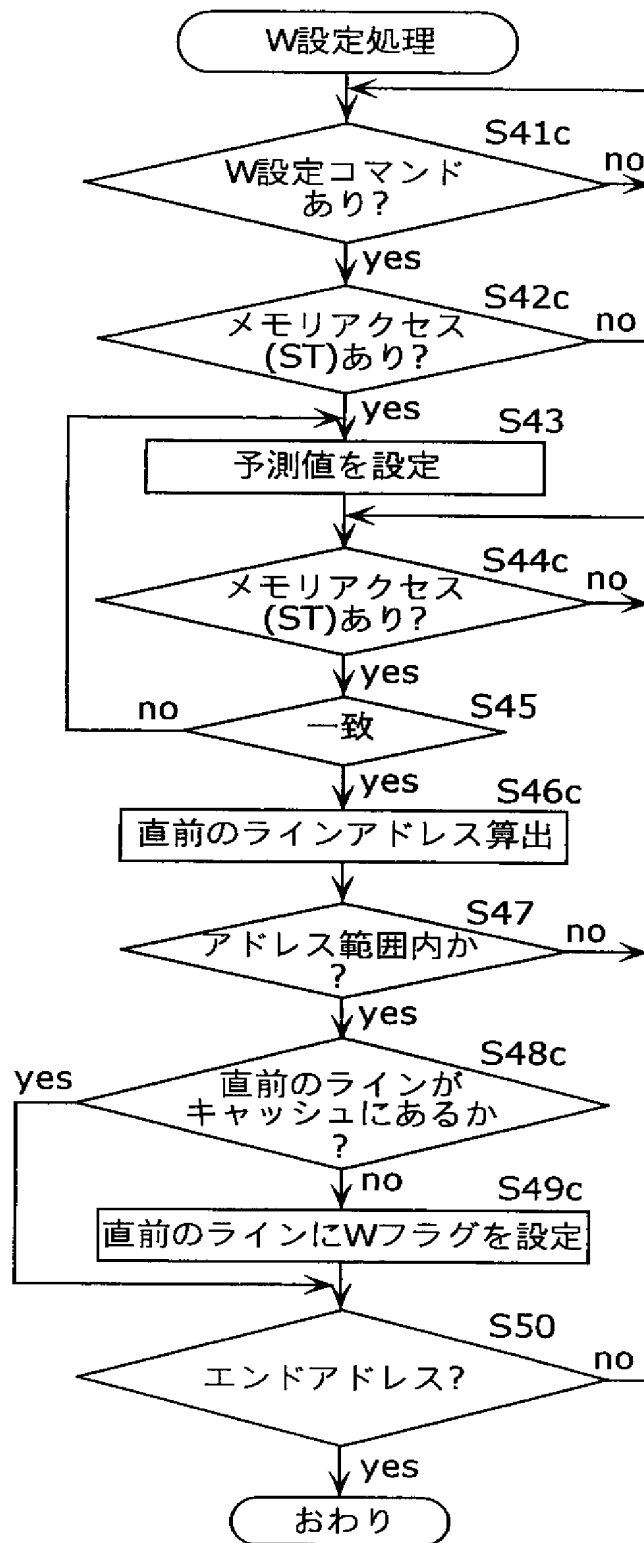
[図14]



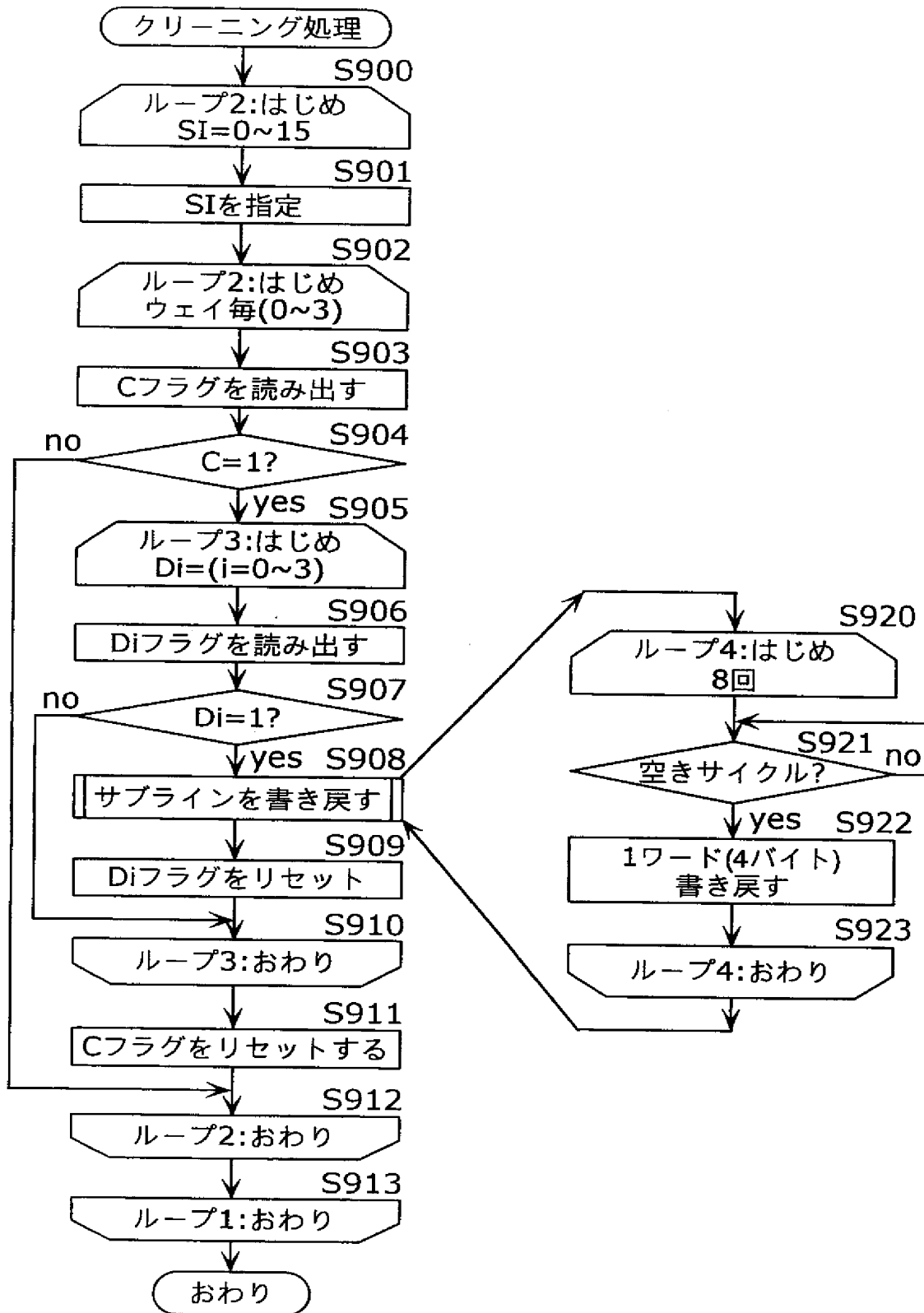
[図15]



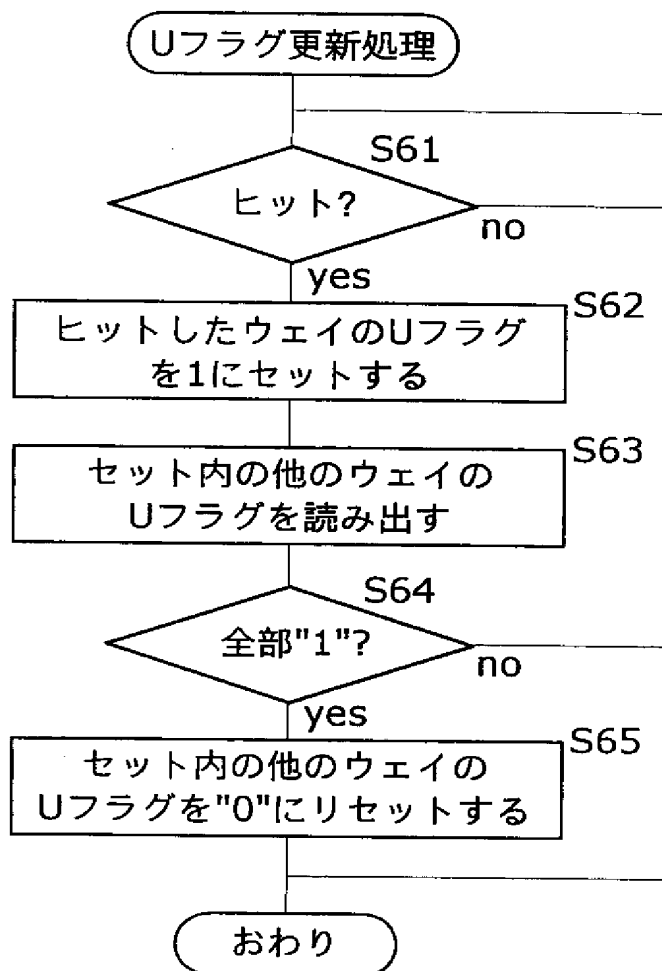
[図16]



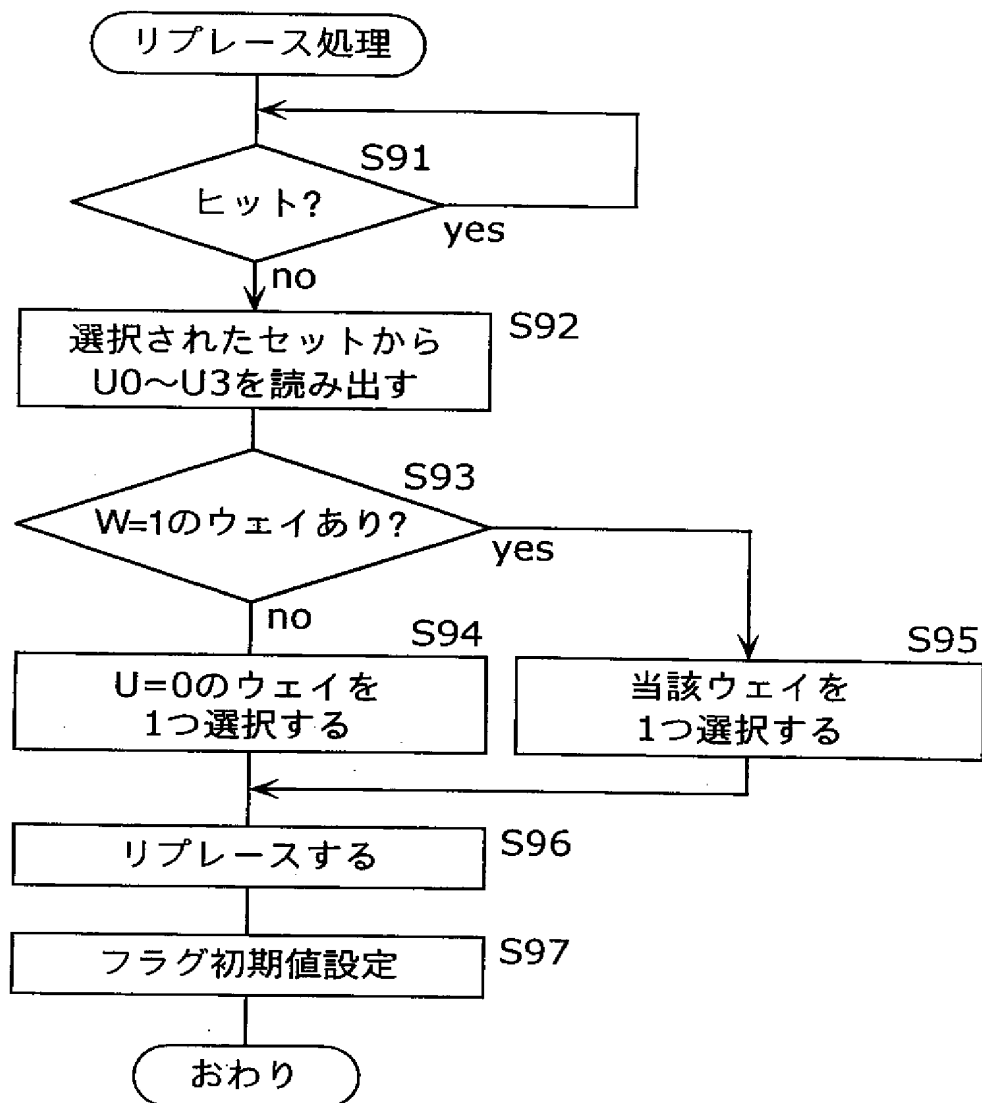
[図17]



[図18]



[図19]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2005/004676

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl.⁷ G06F12/08

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Int.Cl.⁷ G06F12/08-12/12

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2005
Kokai Jitsuyo Shinan Koho	1971-2005	Toroku Jitsuyo Shinan Koho	1994-2005

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP 11-167520 A (NEC Corp.), 22 June, 1999 (22.06.99), Par. Nos. [0021] to [0043]; Figs. 1 to 6	1-4, 6-12, 17, 18
Y	& US 6173392 B1 Column 2, line 22 to column 5, line 41; Figs. 1 to 6C	5, 13, 17
Y	JP 2003-223360 A (Hitachi, Ltd.), 08 August, 2003 (08.08.03), Par. Nos. [0071], [0073] to [0079]; Figs. 6 to 8 (Family: none)	5, 17
Y	JP 7-84879 A (Toshiba Corp.), 31 March, 1995 (31.03.95), (Family: none)	13

☒ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
15 June, 2005 (15.06.05)Date of mailing of the international search report
05 July, 2005 (05.07.05)Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2005/004676

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP 60-45855 A (Fujitsu Ltd.), 12 March, 1985 (12.03.85), (Family: none)	1, 14, 15, 16
A	JP 61-16348 A (Mitsubishi Electric Corp.), 24 January, 1986 (24.01.86), (Family: none)	14, 15
A	JP 51-19453 A (Fujitsu Ltd.), 16 February, 1976 (16.02.76), (Family: none)	16

A. 発明の属する分野の分類 (国際特許分類 (IPC))

Int.Cl.⁷ G06F12/08

B. 調査を行った分野

調査を行った最小限資料 (国際特許分類 (IPC))

Int.Cl.⁷ G06F12/08-12/12

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2005年
日本国実用新案登録公報	1996-2005年
日本国登録実用新案公報	1994-2005年

国際調査で使用した電子データベース (データベースの名称、調査に使用した用語)

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
X	J P 11-167520 A (日本電気株式会社)	1-4, 6-12,
	1999.06.22,	17, 18
Y	【0021】-【0043】、【図1】-【図6】	5, 13, 17
	& US 6173392 B1,	
	第2欄第22行-第5欄第41行、第1図-第6C図	

☒ C欄の続きにも文献が列举されている。☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」 特に関連のある文献ではなく、一般的技術水準を示すもの
「E」 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献 (理由を付す)
「O」 口頭による開示、使用、展示等に言及する文献
「P」 国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」 国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」 同一パテントファミリー文献

国際調査を完了した日

15.06.2005

国際調査報告の発送日

05.7.2005

国際調査機関の名称及びあて先

日本国特許庁 (ISA/JP)

郵便番号100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官 (権限のある職員)

清木 泰

5N

9643

電話番号 03-3581-1101 内線 3586

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
Y	J P 2003-223360 A (株式会社日立製作所) 2003. 08. 08, 【0071】、【0073】－【0079】、【図6】－【図8】 (ファミリーなし)	5, 17
Y	J P 7-84879 A (株式会社東芝) 1995. 03. 31 (ファミリーなし)	13
X	J P 60-45855 A (富士通株式会社) 1985. 03. 12 (ファミリーなし)	1, 14, 15, 16
A	J P 61-16348 A (三菱電機株式会社) 1986. 01. 24 (ファミリーなし)	14, 15
A	J P 51-19453 A (富士通株式会社) 1976. 02. 16 (ファミリーなし)	16